DOCUMENT RESUME

ED 364 581 TM 020 806

TITLE	Proceedings of the CREATE Cross-Cutting Evaluation Theory Planning Seminar (Kalamazoo, Michigan, June 2-3, 1993).
INSTITUTION	Center for Research in Educational Accountability and Teacher Evaluation (CREATE), Kalamazoo, MI.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	[93]
CONTRACT	R117Q00047
NOTE	134p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Accountability; *Educational Planning; Educational Practices; Elementary Secondary Education; *Evaluation Methods; Evaluation Utilization; Integrated Activities; Models; Political Influences; Seminars; *Teacher Evaluation; *Theories
IDENTIFIERS	*Center for Research on Educ Account Teacher Eval

ABSTRACT

The Cross-cutting Evaluation Theory Planning Seminar was initiated to provide insight that will assist in development of a planning proposal for the Cross-cutting Theory Project of the Center for Research on Educational Accountability and Teacher Evaluation (CREATE). CREATE consists of five separate programs, four of which address specific topics in educational accountability and teacher evaluation. The Cross-cutting Theory Project will explore ideas emerging from individual CREATE projects and integrate findings in ways that will facilitate use in schools. Four background papers were prepared as the foundation for the theory development planning seminar. Discussions by the four planning teams followed the background papers. Additional reports from James Stronge, Edward Iwanicki, and Carol Dwyer discuss the evaluation models in more depth, with Michael Scriven providing an analysis of the models. The seminar closed with a summary of the proceedings by William Webster. The four background papers (appendixes A through D) are: (1) "Discussion Draft of a Tentative Work Plan for Project 4.1--CREATE Study Committee on Theory and Practice in Educational Evaluation: Cross-Cutting Theory and Practice" (Daniel L. Stufflebeam); (2) "A CREATE Overview" (Arlen Gullickson); (3) "The Foundations of Educational Accountability" (Michael Scriven); and (4) "Politics of Teacher Evaluation" (Gene V. Glass and Barbara A. Martinez). Six tables illustrate the discussion. (Contains 19 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 364 581

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Daniel L. Stufflebeam

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PROCEEDINGS OF THE CREATE CROSS-CUTTING EVALUATION THEORY PLANNING SEMINAR

June 2-3, 1993

CENTER FOR RESEARCH ON EDUCATIONAL ACCOUNTABILITY AND TEACHER EVALUATION

The Evaluation Center
Western Michigan University
Kalamazoo, Michigan 49008-5178
(616) 387-5895

TM020806

BEST COPY AVAILABLE

For more information contact:

CREATE
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178
Phone: (616) 387-5895
Fax: (616) 387-5923
Internet E-Mail CREATE_Info

This work was supported by the Office of Research, Office of Educational Research and Improvement, U.S. Department of Education (Grant No. R117Q00047). The opinions expressed are those of the authors, and no official support by the U.S. Department of Education is intended or should be inferred.

BEST COPY AVAILABLE

PROCEEDINGS OF THE CREATE CROSS-CUTTING EVALUATION THEORY PLANNING SEMINAR

June 2-3, 1993

CENTER FOR RESEARCH ON EDUCATIONAL ACCOUNTABILITY
AND TEACHER EVALUATION
(CREATE)

THE EVALUATION CENTER
WESTERN MICHIGAN UNIVERSITY

This work was funded by the Office of Educational Research and Improvement, U.S. Department of Education (Grant No. R117Q00047). The opinions expressed are those of the authors, and no official support of these positions by the U.S. Department of Education is intended or should be inferred.

PROCEEDINGS
OF THE
CREATE CROSS-CUTTING EVALUATION THEORY
PLANNING SEMINAR
JUNE 2-3, 1993

CONTENTS

I. Introduction 1

II. Objectives of Seminar 5

III. Seminar Agenda 6

IV. List of Seminar Participants 11

V. Summary of Planning Team Reports 13

REPORT A 13

Planning Team 1 13

Planning Team 2 14

Planning Team 3 15

Planning Team 4 17

REPORT B 19

Planning Team 1 19

Planning Team 2 21

Planning Team 3 23

Planning Team 4 25

VI. Seminar Summary (William Webster) 26

VII. Responses to William Webster's Seminar Summary 31

VIII. Seminar Evaluation Report (Stanley Nyirenda) 38

IX. Appendices

Appendix A: "Discussion Draft of a Tentative Work Plan for Project 4.1 - CREATE Study Committee on Theory and Practice in Educational Evaluation: Cross-Cutting Theory and Practice" (Daniel L. Stufflebeam)

  Attachment 1: Illustrative Project 4.1 Subproject: Development of Sourcebooks for Training Teachers in Educator Evaluation

  Attachment 2: Draft Outline for the Proposed Cross-Cutting Theory and Practice Project

  Attachment 3: Teacher Performance Evaluation System Improvement Kit - A Prospectus

Appendix B: "A CREATE Overview" (Arlen Gullickson)

  Attachment 1: Thumbnail Sketches of Current Development Efforts

Appendix C: "The Foundations of Educational Accountability" (Michael Scriven)

Appendix D: "Politics of Teacher Evaluation" (Gene V Glass and Barbara A. Martinez)

CREATE CROSS-CUTTING EVALUATION THEORY PLANNING SEMINAR

The CREATE Cross-cutting Evaluation Theory Planning Seminar was initiated to provide insight that will assist in the development of a planning proposal for the final two years of CREATE's Cross-cutting Theory Project.

BACKGROUND

In order to fulfill its mission to improve educational evaluation in U.S. schools, the Center for Research on Educational Accountability and Teacher Evaluation (CREATE) must improve the underlying theory that drives evaluation practices. CREATE projects individually meet this expectation, but much good information can be overlooked or otherwise lost in the process. The Cross-cutting Theory Project was initiated to provide cross-project integration, permitting what could be called an "interaction effect." That is, if CREATE ideas are discussed in a cross-project and cross-program context, new implications and applications will be permitted to emerge.

CREATE consists of five separate programs, four of which address specific topic areas in educational accountability and teacher evaluation, and 15 projects in total with 11 currently existing or planned for years 4 and 5. Projects have been grouped by topic to facilitate exchange of ideas and information across projects. However, each project has its own objectives, agenda, and product requirements.

The CREATE programs and associated (past and present) CREATE and Evaluation Center projects on which this integration effort draws are as follows (project staff are listed in parentheses with dates of project):

Program 1: Improvement of Teacher Performance Evaluation

- *Teacher Evaluation Models Project (Scriven/Wheeler/Haertel: 11/90 - 10/93)*
- *Improved Teacher Evaluation Models Development (Scriven: 11/93 - 10/95)*
- *Grounded Theory of Teacher Performance Evaluation (Stufflebeam/Thomas: 1/92 - 10/93)*
- *Development of Classroom Assessment Techniques for Teacher Self-Evaluation (Gullickson/Airasian: 5/92 - 10/94)*
- *Expert Science Teacher Project (Burry: 11/90 - 10/93)*
- *National Fast Response Survey of Teachers (Stufflebeam: 12/91 - 12/93; Funded by NCES, not directly a CREATE project)*
- *National Survey of Schools (Barley: 11/90 - 10/91)*

Program 2: Improvement of Administrator and Support Personnel Evaluation

- *Improvement of Administrator and Support Personnel Evaluation Theory Development and Special Projects*

- *Models for Administrator Evaluation (Stufflebeam/Bridges/Candoli/Cullen: 1/93 - 10/95)*
- *Model for Evaluation of Professional Support Personnel (Stronge/Helm: 11/92 - 10/95)*

Program 3: Improvement of School Evaluation

- *Improved and Tested School Evaluation Models (Sanders: 11/93 - 10/95)*
- *Development of a Research-Based School Report Card (Nyirenda/Jaeger: 05/92 - 10/93)*
- *Models for School Evaluation (Gallegos/Benjamin/Candoli/Wegenke: 11/90 - 10/92)*

Program 4: Theory Development and Special Projects

- *CREATE Study Committee on Theory and Practice in Educational Evaluation: Cross-cutting Issues and Practice (Stufflebeam/Ervin/Gullickson: 11/92 - 10/95)*
- *Adapting W. Edwards Deming's Statistical Process Control Model (Bayless/Massara: 1/91 - 10/92)*
- *Evaluating the Ability to Work With At-Risk Students and Their Families (Lavely/Blackman: 11/90 - 10/92)*

Program 5: National Evaluation Resource Service

- *Dissemination of Research and Products (Gullickson/Ervin: 11/90 - 10/95)*
- *National Evaluation Workshops (Gullickson/Sandberg: 11/90 - 10/92)*
- *Evaluations of Nationally Significant Evaluation Programs: (11/90 - 10/92)*

The Cross-cutting Theory Project will explore ideas that emerge from individual CREATE projects, identify consistencies and commonalities across projects, resolve cross-project discrepancies, integrate findings across projects, and combine and package CREATE products in ways that will facilitate use in schools. The project will provide a forum for synthesis of project efforts, a place to expose problems, to understand trends, to provide solution strategies that cross project and program boundaries, and to design products for use by practitioners. This project's principal products will be position papers, monographs, books, and training materials deriving from CREATE's ongoing investigations. These products will develop overarching theory, identify fallacies and weaknesses in present findings, modify individual project findings and directions, and provide guidelines for practice. This process will provide an "interaction effect," ensuring that the whole of CREATE is greater than the sum of its parts. Through this integrative approach the project will help CREATE develop

(a) a more comprehensive theory for understanding and guiding evaluation practice

(b) a cohesive set of supporting materials to guide educators in their conduct of personnel and school evaluation practices and in preparing educators to be proficient in conducting and using evaluations

(c) sound ways of addressing issues that cut across CREATE's research and development

THE PLANNING SEMINAR

During the first year of the Cross-cutting Theory Project, four background papers were prepared as the foundation for the theory development planning seminar that was held in June 1993. Participants in the seminar included project staff, the CREATE National Advisory Panel, and invited scholars and practitioners. Many Advisory Panel members are involved directly in CREATE projects as well as in policy direction and oversight. This format made it possible for Panel members and project directors to have substantive input and was intended not only to serve the development of integrative approaches to evaluation, but also to provide direct feedback to individual project directors and staffs.

The four background papers presented at the seminar were an analysis of CREATE projects, written by Dr. Arlen Gullickson (Appendix B); a conceptual paper developed by Dr. Michael Scriven (Appendix C), which proposes an overall structure for use in arranging and integrating CREATE contributions toward identifying pervasive themes across the subdisciplines in evaluation work; a paper by Dr. Gene Glass that focuses on political issues and paradoxes that must be considered in developing a functional theory of educational evaluation (Appendix D); and a draft plan for the 1993-95 project work by Dr. Daniel Stufflebeam (Appendix A).

The seminar began with the presentation of these papers, followed by group discussion among four planning teams. The team participants were selected to ensure that each team had representatives from the Advisory Panel, project directors, invited guests, and CREATE staff. An attempt was made to further ensure that both researchers and practitioners were represented on each team.

On the first day, the teams were given five questions to address relating to the direction the Cross-cutting Theory Project should take with regard to prioritizing audiences, issues, and outcomes. The second day was to be devoted to team discussion of the most important focus for the project, followed by team reports. However, on the second morning of the seminar, a change in direction occurred. At the request of the participants, the models were discussed in more depth. Reports were heard from Dr. James Stronge, Dr. Edward Iwanicki, and Dr. Carol Dwyer describing their respective models, followed by an analysis of the models by Dr. Michael Scriven. This discussion sparked interest in the potential for one "common model" of evaluation that could be applied to teachers, administrators, and professional support personnel. The agenda attached is revised to reflect what occurred at the seminar.

The seminar closed with a summary of the proceedings by Dr. William Webster, followed by responses from the audience.

The reports from the Planning Teams are presented as Report A (response to the five questions posed to the teams) and Report B (suggestions for the most important focus of the project).

These reports, along with the papers and summary presented in these proceedings, will provide guidance to the Cross-cutting Theory Project staff as they plan the direction of the project in 1994 and 1995.

OBJECTIVES OF SEMINAR

The objectives of the CREATE Cross-cutting Evaluation Theory Planning Seminar were

A. To inventory the lessons from CREATE's projects, to begin organizing these lessons conceptually, to examine the potential contributions of CREATE's projects to improving school evaluation systems, and to delineate study questions and development objectives that cut across projects

B. To identify issues related to the politics of evaluation and to consider how CREATE should address the issues

C. To help CREATE develop a shared view of the requirements of sound theory and determine what steps it should take vis-a-vis the development of evaluation theory

D. To help focus and develop CREATE's plan for combining and disseminating findings and products from CREATE's projects in order to address cross-cutting issues and improve evaluation practices in schools

SEMINAR AGENDA

DAY 1

Wednesday, June 2

Time	Activity	
8:30 am	Overview of the seminar objectives and introductions -- 20 min.	Daniel Stufflebeam
8:50 am	Overview and analysis of approaches and findings from CREATE's projects, followed by discussion "A CREATE Overview" -- 80 min	Arlen Gullickson
10:25 am	Developing and Using Theory to Improve Evaluation Practice, followed by discussion from the floor "The Foundations of Educational Accountability" -- 75 min.	Michael Scriven
1:10 pm	"Politics of Teacher Evaluation," followed by reactions from three selected participants and general group discussion -- 75 min	Gene Glass
2:25 pm	Initial plan for combining findings and products from CREATE's projects in order to address cross-cutting issues and develop practical tools and strategies for use in improving school evaluation practices, followed by discussion from the floor "Discussion Draft of a Tentative Work Plan for Project 4.1" -- 60 min.	Daniel Stufflebeam
3:40 pm	Organization and charge to planning teams -- 20 min.	Arlen Gullickson
4:00 pm	Meetings of 4 planning teams to address the following questions:	

1. **What relative priorities should this project assign to serving different constituent groups, especially evaluation researchers, school practitioners, and teacher educators?**

2. **What cross-cutting issues should receive this project's highest priorities, e.g., the politics of evaluation, possibilities of integrating personnel and program evaluation, use of student performance measures to evaluate teachers, achieving sound teacher evaluation in the context of collective bargaining agreements?**

3. Should this project put its emphasis on periodic cross-cutting issues seminars, concentrate on product development efforts, do a combination of these, or do something else?

4. What outcomes should this project seek, e.g., issue papers, materials for training teachers in the principles and methods of evaluation; strategies and tools for schools to use in analyzing and improving their teacher evaluation systems, etc.?

5. What would be the most important issue to pursue during this seminar?

-- 90 min.

7:00 pm Planning teams continue their work -- 90 min.

8:30 pm Small group reports, response by Gullickson and Stufflebeam, and discussion from the floor -- 45 min.

9:15 pm Adjourn

DAY 2

Thursday, June 3

Time	Activity	
8:00 am	New charge to planning teams: What is the most important focus for the Cross-cutting Theory Project? Planning teams address their assignments and develop their reports -- 90 min.	Arlen Gullickson Daniel Stufflebeam
9:30 am	Reports by planning teams (spokesperson for each team) -- 30 min.	
10:00 am	Presentation of models followed by analysis of models -- 120 min.	James Stronge Edward Iwanicki Carol Dwyer Michael Scriven
1:30 am	Seminar Summary -- 55 min.	William Webster
2:25 pm	Group discussion -- 30 min.	Daniel Stufflebeam
2:55 pm	Wrap-up and overview of next day's activities -- 20 min.	Arlen Gullickson Daniel Stufflebeam

GROUP AND INDIVIDUAL ASSIGNMENTS

Seminar Manager	Arlen Gullickson
Seminar Logistics and Reporter	Edith Ervin
Seminar Evaluator	Stanley Nyirenda
Seminar Secretary	Sally Veeder
Travel Arrangements	Patti Negrevski
CREATE Director	Daniel Stufflebeam
Resource Persons on CREATE findings	Arlen Gullickson
	Project Directors
Resource Person on Politics of Evaluation	Gene Glass
Resource Person on Theory Development	Michael Scriven
Resource Person on Cross-cutting plan	Daniel Stufflebeam
Panel to React to the Glass Paper	Carol Norman
	Edward Iwanicki
	Phil Robinson
Planning Team 1	Benjamin Alvarez
	Gilbert Austin
	Edith Ervin (Summarizer)
	Ruben Olivarez
	James Sanders
	Anthony Shinkfield (Chair)
Planning Team 2	Peter Airasian
	Candis Baker
	Carol Dwyer (Chair)
	Gene Glass
	Phil Robinson
	Rebecca Thomas (Summarizer)
Planning Team 3	Bruce Gould (Chair)
	Conrad Katzenmeyer
	Robert Rodosky
	Michael Scriven
	Sally Veeder (Summarizer)
	Gary Wegenke

Planning Team 4	Edward Iwanicki Jason Millman Carol Norman (Chair) Mary Gourley (Summarizer) James Stronge Robert Ward
Floaters (across teams)	Arlen Gullickson Stanley Nyirenda Daniel Stufflebeam William Webster

LIST OF SEMINAR PARTICIPANTS

Invited Guests

Dr. Benjamin Alvarez
International Youth Foundation
67 W. Michigan Ave., Suite 608
Battle Creek, MI 49017

Dr. Carol Dwyer
Educational Testing Service
Princeton, NJ 08541

Dr. Edward F. Iwanicki
Department of Educational
Leadership U-93
School of Education
University of Connecticut
Storrs, CT 06268

Dr. Conrad Katzenmeyer
National Science Foundation
Division of Research, Evaluation,
and Dissemination
1800 G Street, Room 1249
Washington, DC 20550

Dr. Ruben Olivarez
Texas Education Agency
1701 N. Congress
Austin, TX 78701-1494

Dr. Phil C. Robinson
PR-2 and Associates
1367 Joliet
Detroit, MI 48207

Dr. Robert Rodosky
Director of Research
Jefferson County Public Schools
VanHoose Educational Center
3332 Newburg Road
Louisville, KY 40218

Dr. William J. Webster
Dallas Independent School District
3700 Ross Avenue
Dallas, TX 75204-5491

CREATE National Advisory Panel

Dr. Gilbert Austin, Director
Center for Educational Research
and Development
University of Maryland
Baltimore, MD 21228

Ms. Candis Baker, Chair
Science Department
Milwood Middle School
2916 Konkle
Kalamazoo, MI 49001

Dr. Gene V Glass
Division of Educational Leadership
and Policy Studies
College of Education
Arizona State University
Tempe, AZ 85287-2411

Dr. Bruce Gould, Senior Scientist
AL/HRMM
Brooks AFB, TX 78235-5601

Dr. Jason Millman, Professor
Educational Research Methodology
Department of Education
405 Kennedy Hall
Cornell University
Ithaca, NY 14853-5901

Dr. Carol Norman, Manager
Research, 6th Floor
National Education Association
1201 16th Street, N.W.
Washington, DC 20036-3290

Dr. Anthony Shinkfield, Evaluator
49 Mann Terrace
North Adelaide
South Australia

Mr. Robert Ward, Principal
Lakeview High School
300 South 28th Street
Battle Creek, MI 49015

Dr. Gary L. Wegenke, Superintendent
Des Moines Public Schools
1800 Grand Avenue
Des Moines, IA 50307-3382

CREATE Staff

Dr. Peter Airasian, Project Consultant
Boston College
College of Education
Campion 336D
Chestnut Hill, MA 02167-3813

Ms. Edie Ervin
Director of Project Information
and Product Development
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Ms. Mary Gourley
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Dr. Arlen Gullickson
Associate Director-CREATE
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Dr. Stanley Nyirenda
Internal Evaluator-CREATE
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Dr. James Sanders
Evaluation Education Specialist
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Dr. Michael Scriven, Project Director
P.O. Box 69
Point Reyes, CA 94956

Dr. James H. Stronge, Project Director
School of Education
The College of William and Mary
Williamsburg, VA 23185

Dr. Daniel L. Stufflebeam, Director
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

Ms. Rebecca Thomas
Assistant to the Director
Western Michigan University
Kalamazoo, MI 49008-5178

Ms. Sally Veeder, Assistant Director
The Evaluation Center
Western Michigan University
Kalamazoo, MI 49008-5178

SUMMARY OF PLANNING TEAM REPORTS

Each team developed two reports. ***Report A*** summarizes the team's responses to the five questions posed to participants. Some teams answered the questions individually and some chose to brainstorm on specific projects, such as a teacher evaluation kit for schools.

Report B summarizes what the team viewed as the most important focus for the cross-cutting theory project.

These reports were generated to stimulate ideas and provide additional insights and suggestions and should be viewed as discussion documents rather than cohesive plans.

REPORT A: QUESTIONS POSED TO PLANNING TEAMS

1. What relative priorities should this project assign to serving different constituent groups, especially evaluation researchers, school practitioners, and teacher educators?

2. What cross-cutting issues should receive this project's highest priorities, e.g., the politics of evaluation, possibilities of integrating personnel and program evaluation, use of student performance measures to evaluate teachers, achieving sound teacher evaluation in the context of collective bargaining agreements?

3. Should this project put its emphasis on periodic cross-cutting issues seminars, concentrate on product development efforts, do a combination of these, or do something else?

4. What outcomes should this project seek, e.g., issue papers, materials for training teachers in the principles and methods of evaluation; strategies and tools for schools to use in analyzing and improving their teacher evaluation systems, etc.?

5. What would be the most important issue to pursue during this seminar?

Planning Team 1

Benjamin Alvarez
Gilbert Austin
Edith Ervin (Summarizer)
Ruben Olivarez
James Sanders
Anthony Shinkfield (Chair)

Responses to the questions:

1. Emphasis should be placed on school practitioners and then teacher educators.

2. Issues should reside at the school level. The group had some concerns about integrating personnel and program evaluation. CREATE should provide guidelines for a holistic approach--integrating aspects of various evaluations. There should be a clear delineation of duties at all levels and a clear alignment of these. A kit should be developed by CREATE that is sensitive to various levels [e.g. teachers, administrators] for both school and personnel evaluation. Such a kit should include critical questions, essential role of the principal, and common criteria for evaluation. Notice should be taken both of information available from effective schools as well as the importance of student performance measures.

3. The group recommends equal emphasis on seminars and product development--one feeds on the other and both help to identify agendas.

4. The priority should be strategies for schools to develop methods of improvement of teacher evaluation systems. Guidelines should be interpreted by schools based on collaborative efforts of all groups connected with the school. The project should also develop a training package for preservice and inservice education in evaluation.

5. Questions to be pursued during this seminar:

    5.1 What is a good teacher evaluation and how can a staff development paradigm help to explore this issue?

    5.2 How do we assess the effectiveness of a school?

    5.3 How can a comprehensive evaluation of a school be undertaken encompassing all essential elements? Take note that what would be included would be definition of components, role responsibilities, and what they are and how they relate.

Planning Team 2

Peter Airasian
Candis Baker
Carol Dwyer (Chair)
Gene Glass
Phil Robinson
Rebecca Thomas (Summarizer)

Responses to the questions:

1. Highest priority should be given to focus on service to professional teachers regardless of who does the evaluation, administrator or teacher.

2. Cross-cutting issues that should receive this project's highest priories:
   Equity/fairness (inclusion/exclusion)
   Training assessors
   Power/status
   Preservice piece

3. This project should put its emphasis on the development of products to be used for direct inservice efforts by teachers.

4. Development of inservice materials should include topics of emerging interest to teachers and topics dealing with teachers' rights in evaluation.

5. Questions to be addressed deal with the themes of equity and fairness cutting across all of the projects in CREATE.

Planning Team 3

Bruce Gould (Chair)
Conrad Katzenmeyer
Robert Rodosky
Michael Scriven
Sally Veeder (Summarizer)
Gary Wegenke

Responses to the questions:

1. Relative priorities assigned to constituent groups:

   - Evaluation researchers and teacher educators

   - Policymakers at federal level (legislators and people in agencies). This group is missing from named constituent list.

   Need to create interest to sensitize consumers to value/need for evaluation

2. Cross-cutting issues to receive priority:

   All those mentioned (the politics of evaluation, possibilities of integrating personnel and program evaluation, use of student performance measures to evaluate teachers, achieving sound teacher evaluation in the context of collective bargaining agreements) are a part of a total system.

CREATE could produce a monograph on integration of program and personnel evaluation.

Some attention should be paid to the use of student performance measures to evaluation of teachers.

We should establish a system of standards and procedures for addressing all these issues at three levels:

- Professional standards
- Duty
- Coal-face (content level): a process for saying the process is good enough or not good enough

3. What should this project emphasize?

The CREATE newsletter should be used to disseminate cross-cutting issues by summarizing potential applications of the cross-cutting work (short, succinct, user-friendly blurbs).

Someone (a constituent or representative of CREATE) should go to DC to talk with policymakers.

NSBA would be a good organization to work with. It controls a vast majority of the money in education.

The seminars are expensive, but a good value. In choosing between the seminars or products, we chose a combination (do both).

4. Outcomes to be sought:

Materials for trainers

"Gem" products/projects need to be brought to the attention of school professionals.

Provide draft legislation to legislators. We are not sure that this is in CREATE's charter (proposing legislation), since CREATE's mission is focused on research and development. However, probably the greatest service we could do to be sure evaluation is institutionalized is to sensitize legislators and provide them with proposed legislation.

5. What should be discussed at seminar?

   What is CREATE doing that would be of use to school administrators?

   Is there a common evaluation model to be used across the board?

   Is James Stronge's evaluation model usable with all groups: teachers, administrators, support personnel?

   The Personnel Evaluation Standards are metastandards; Scriven's duties cover another element. Is this notion correct or not?

   How should the school evaluation model involve administrator evaluation and teacher evaluation?

   How should the teacher evaluation model involve school and administrator evaluation?

   How should the administrator evaluation model involve school and teacher evaluation?

Proposed Project:

Adopt a school system. Ask personnel in that system the four questions given to work groups, and write up the results as a case study.

Planning Team 4

Edward Iwanicki
Jason Millman
Carol Norman (Chair)
Mary Gourley (Summarizer)
James Stronge
Robert Ward

Brainstorm - List of Kit Components

1. Teacher and evaluator Bill of Rights (equity, fairness, legal issues)
2. Communication-Conferencing skills, etc.
3. Critical questions to get at evaluation purposes and context
4. Training of evaluators
5. Use of student ratings

6. Use of student performance
7. Teacher evaluation/collective bargaining
8. Resources for teacher evaluation
9. Building collaboration groups for evaluation - role of colleagues in teacher evaluation
10. Role descriptions - responsibilities
11. Objective performance measures
12. Judgmental input by experts
13. Measurement of subject matter knowledge
14. Classroom management skills
15. Assessment skills
16. Aggregating information for decision making
17. Personnel evaluation standards applied to teacher evaluation
18. Pitfalls in teacher evaluation

Items from Kit Paper - Stufflebeam

REPORT B: WHAT IS THE MOST IMPORTANT FOCUS FOR THE CROSS-CUTTING THEORY PROJECT?

Planning Team 1

Benjamin Alvarez
Gilbert Austin
Edith Ervin (Summarizer)
Ruben Olivarez
James Sanders
Anthony Shinkfield (Chair)

Topic: How can the evaluation of teachers be related to the evaluation of their school and its administration?

The group emphasized that any outcome of the cross-cutting project should focus on accountability.

1. CREATE should use extant material available in research and other literature on successful schools and successful teachers.

2. Comprehensive evaluation essentially involves critical questions and definitions of critical characteristics. These are then juxtapositioned against a particular school.

    2.1 If CREATE offered a model for context evaluation, these critical questions and characteristics could be addressed. Not escaping the net would be school boards, who should be included at least conceptually [and finance would dictate whether the comprehensive evaluation would include such a group].

    2.2 A context evaluation must include defined roles and responsibilities.

3. Accountability must be discussed up front in any kit or manual produced by CREATE in this context. Such placement is a form of moral obligation to the school and its community. For example, the school board's defined integrity in the complete system is seen as an important element of the accountability statement.

    3.1 Similarly, all personnel connected with the school must be shown to be accountable and the reasons given.

    3.2 Accountability is attaching responsibility to school evaluation. Inherent is the development of a set of expectations--clearly outlined--about the use of valuable resources. There should be rewards and also appropriate actions for failure to use these resources well.

3.3 The principal as the instructional leader must be both knowledgeable and effective in teacher evaluation.

3.4 Related to accountability, there should be common indicators based on a range of performance outcomes [of which student assessments are simply one].

3.5 Accountability measures must be developed to lead toward improvement in pedagogy and, most importantly, student learning.

4. Standards and procedures for evaluation

Formative evaluation should lead as naturally as possible into summative evaluation based on collaboratively accepted criteria for judgment. Such indicators must include a wide range of methods and approaches including duties, proven successful techniques, as well as many forms of [testable] student performance.

4.1 In general terms, the various groups associated with the school should be evaluated using the same standards.

4.2 The school evaluation component similarly should include a wide range of approaches. The work being undertaken by CREATE in this regard should be directly applicable.

5. What are CREATE's responsibilities?

5.1 Summarize characteristics of effective schools and duties of teachers [Scriven].

5.2 List characteristics and other relevant information for inclusion in a kit [refer to critical questions mentioned earlier].

5.3 As mentioned earlier, design context evaluation.

5.4 Include resources that are part of effective teaching, as for example, the new technologies and teachers' role in using these.

5.5 Concerning school evaluation, pool together provisions achieved to date, e.g., consumer report, school report cards, and literature identifying strengths and weaknesses of existing models. Ultimately develop one or more models for field testing--followed by the production of a manual or kit [as part of the total kit, perhaps, for the kind of comprehensive evaluation referred to in this report].

6. The connection between teacher and school evaluation

6.1 In any discussion about the quality of a school, one vital dimension is the quality of teaching. Thus, the evaluation of one cannot be separated from that of the other.

6.2 Both must be consonant--use what we already have discovered from work carried out by CREATE.

6.3 In both kinds of evaluation, criteria definition and role explication are basic, as are definitions of the functions of a school--including, importantly, the basic mission of delivering the best curriculum.

7. Budget

7.1 If this proposal is accepted, it would become an integral part of future seminars--and thus difficult to separate a cost component.

7.2 It is estimated that the comprehensive school evaluation kit would cost $40,000 for each of two years.

8. Appendix A

The group offered unequivocal and strong support for a project leading toward preservice and inservice training of educators in evaluation.

Planning Team 2

Peter Airasian
Candis Baker
Carol Dwyer (Chair)
Gene Glass
Phil Robinson
Rebecca Thomas (Summarizer)

A more concise question to address:
What might be in the kit that is oriented toward fairness?

OVERARCHING THEME FOR ALL CREATE MATERIALS: IMPROVING PRACTICE

Central validity: Use of the Personnel Evaluation Standards
Adoption of the Standards as a means of avoiding "bad things" that can occur in the systems of personnel evaluation.

All CREATE materials should focus on feedback that provides direction to teachers; we do not need another set of theoretical materials. Based upon the theory and research, materials must be useful and practical for use of teachers.

PURPOSE OF THE KIT

For inservice education develop:

- Interactive materials
  - Examples
  - Case studies

Materials sensitive to meeting the needs of the children served: Reference Carolyn Lavely & Joe Blackman work on at-risk students.

NATURE OF MATERIALS

The KIT should present practical examples, not as a theoretical series of ethical principles. These examples should help in specific examples of equitable process, inequitable process, fair process, and unfair process (reference PRAXIS Guidelines for Practice materials for these kinds of examples).

Use of language in the KIT should not build stereotypes often underlying educational materials.

TYPES OF MATERIALS (ethical considerations)

Include in KIT as separate pullouts:

Reference Strike's work on Teacher Bill of Rights

Bill of responsibilities for teachers toward kids

Bill of responsibilities for the evaluator toward the teacher

Models should be developed that are not rigid, but flexible enough to be useful to their formative and summative assessment. These should include a focus on teachers taking charge of the process.

Use the theory to connect activities to teachers and classroom problems; active exercises as well as expository materials.

PROCESS FOR DEVELOPMENT OF KIT

Consider a partnership with a district, with an existing interest and compatible philosophy in evaluation from the point of view of the teacher.

TITLES FOR MATERIALS

"WHAT CAN WE DO TO IMPROVE EVALUATION"

This is to develop a sense of personal responsibility teachers must take for their own evaluation and the evaluation process.

Planning Team 3

Bruce Gould, Chair
Conrad Katzenmeyer
Robert Rodosky
Michael Scriven
Gary Wegenke
Sally Veeder, Summarizer

The cross-cutting project should focus/integrate on what should be taken into account by administrators. We need to ask ourselves what CREATE is doing/producing that would be of interest to school administrators. The superintendent of a well-functioning school district may not be interested in CREATE's work or findings. However, even in well-functioning districts, there are areas for improvement and threats of possible legal action.

The issue is how to approach a school district with CREATE project results.

In well-functioning school districts:

- Squeaky minority group (small constituency)

  Thinks that something is wrong
  Desires to improve
  Focuses on an area
  Threat of legal trouble

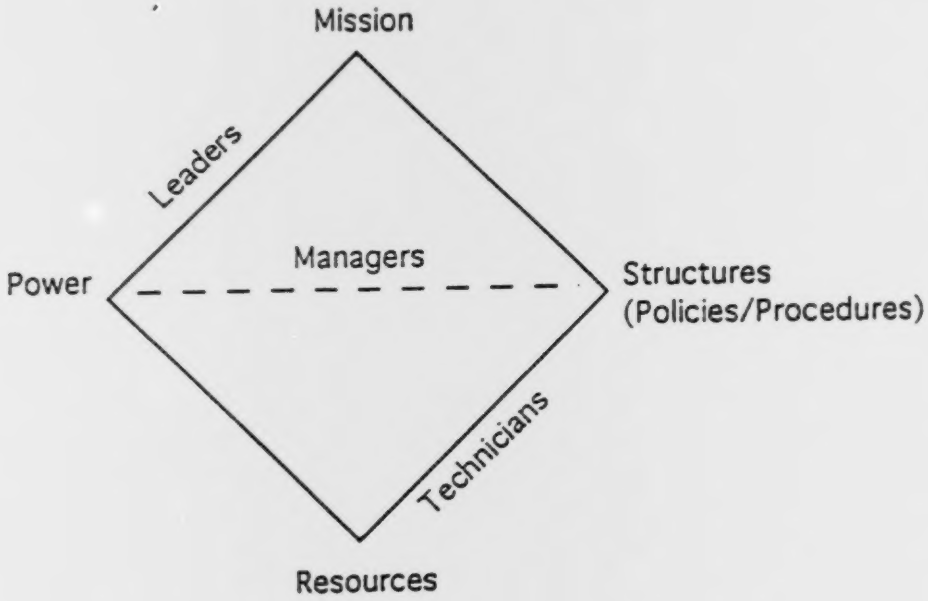
New or nonfunctioning school district:

- Where would evaluation skills be focused?

Organization as a whole--e.g., where is the administration focusing its attention? (organizational structure and resources vs. the more productive mission emphasis and empowering those responsible with ability to meet their mission components?)

- Teaching and learning (curriculum)
- Professional development

This group was impressed with the Robert Terry strategic planning and thinking model used in Des Moines and has reproduced it here:



Leaders operate between Power and Mission. Technicians, who operate between Structures and Resources, limit the organizational development. Managers operate between Power and Structures.

RESPONSE TO SOURCEBOOK PROPOSAL

The group felt this proposal was a good idea, but wasn't sure about its priority level, given the $44,000 budget per year, within a total budget limitation of $220,000. CREATE might consider the possibility of authors working for free and receiving the royalties (DLS suggestion). If it is done, it is to start with a focus on incentives for inservice. Those incentives should focus on both personal and professional development of the teachers. The group was not sure, for the preservice, how much room there is in curricula. Is it worth giving it priority, because colleges of education may not use it?

A critical issue is that before such a sourcebook is developed, the teacher evaluation models must be pulled together into one operationally feasible model. There are at least five candidate models.

A first attempt should look for the commonalities and differences in those five models (this took place at the cross-cutting theory conference). Use that experience to identify whether it is feasible to make this a primary topic for the next cross-cutting theory conference and, if so, to suggest approaches. If there is such a cross-cutting conference, we propose the following agenda:

Day 1: Researchers pull together the common components and identify components of disagreement (those disagreements could form a research agenda).

Day 2: Researchers present the Day 1 results to the administrators/school professionals.

We hope that the administrators/school professionals could help put the model into succinct, practical terms that would be understandable and usable by the target teachers.

Planning Team 4

Edward Iwanicki
Jason Millman
Carol Norman (Chair)
Mary Gourley (Summarizer)
James Stronge
Robert Ward

Issues to address:

1. Resources for training and implementation
2. Theoretical/conceptual underpinnings
    - identify common and conflicting elements across projects
    - work toward the development of a generalizable evaluation model
3. Role of staff development in integrating personnel evaluation and school improvement
4. Submit CREATE work to scrutiny

    Using Personnel Evaluation Standards with particular emphasis on feasibility

SEMINAR SUMMARY

William Webster
Dallas Independent School District

When I first prepared these remarks, I really prepared most of them last night, which of course was before the conference took a major turn this morning. I might add that I think the major turn was for the good. I think the last two or two-and-one-half hours of discussion were extremely enlightening and productive, and I believe that before a lot of the other things I am going to be talking about can be successfully implemented, we probably need about two more days of discussion to try and iron out the models.

From talking to all of you, looking at the products, and listening, I've reached the conclusion that CREATE needs to reach three distinct audiences with its materials. The first, and the one that has probably been talked about the most here, are the practitioners; teachers and administrators working in the schools. The second, and probably equally important from the standpoint of the credibility of the project, is the evaluation research community; and, the third is policymakers. I think, in terms of the policymaker issue, the major failing of our profession has been our inability, or unwillingness, to deal with policymakers. Evaluation comes and goes. As you know, research funding comes and goes, because we have never been able to really get a consistent amount of support for evaluation from the people who are actually controlling the purse strings, those actually making the policies. I think a lot of our success in Dallas has been due to the fact that we have been able to get backing from our policymakers, so that no matter who the superintendent is and what the administration is, the policymakers are demanding accountability. I also think, in terms of these three different audiences, that the materials that you develop need to be differentially developed for each audience. In other words, even if they might be dealing with the same types of topics, they need to be different types of materials for each audience.

For the practitioners, who are the administrators, teachers, and what have you, I think what is required are simple, straightforward solutions to complex problems. That is not to say that it is a simple solution, straightforward explanation of ways to implement the systems, etc. I think Dan's idea of kits is an excellent idea. Educators normally, because they have many other things to do and have many other agendas, are not going to sit down and read a 30-page document. It has been my experience in Dallas that, when documents exceed 10 pages, you lose your audience, just as presentations which exceed 10 minutes lose their audience. So it needs to be simple, straightforward, hard-hitting; but that is not to say that it doesn't have to be intellectually correct, and that is why I think that you need, or we need, or whoever is involved needs, two or three more days of really sorting out the concepts that we are going to attempt to examine.

From the standpoint of the evaluation/research community, I think we need scholarly papers or scholarly pieces on the research behind the solutions to the complex problems mentioned under the practitioners. In other words, reviews of the literature. I think Dan and Carol did a review of literature that is approximately 75 pages long. That is excellent for the research community; but I think, in terms of practitioners, they wouldn't get past the first 5 pages. In other words, you've got to get it more hard-hitting and more to the point. Also, I think that it is very important that you provide the papers or the articles to support the practitioners' solutions, and that can be done through professional papers and what have you. I noted that most of things that Arlen has in Table 7 of his paper really, in my opinion, are not items that would be particularly appropriate for practitioners. They are more appropriate for the second level, for researchers, evaluators, people interested in the problems.

And, finally, in terms of policymakers, I think you need to make a major effort to position the materials in such a way that all understand that the implementation of these systems can bring, in

Michael's terms, demonstrable and creditable responsibility to the education profession. In other words, accountability. Essentially, policymakers are interested in accountability at this point in time. I'm going to elaborate a little more on each of these areas.

In terms of practitioners, I think the major concentration needs to be on personnel evaluation systems. When I talk about personnel evaluation systems, I'm talking about integrating what you call school report cards with administrator evaluation, with teacher evaluation; in other words, school effectiveness indices, if in fact you take that approach, working down to administrator effectiveness indexes, and working down to teacher effectiveness indices. I think this is an excellent time to do that, because I think now, as much as any time in history, most practitioners recognize that current systems are inadequate. I think that's a generally safe statement no matter if they have one observation a year, two observations a year, or five observations a year. In most cases, the things that are being observed, and I can speak specifically for Dallas and for the state of Texas, are not necessarily the kinds of things that the people think are terribly valid. The necessary solution (and maybe I was overreacting to one of the papers), for it seemed to me that one of the solutions that was being suggested was to turn the design of the system over to teachers, is that all stakeholders in the process need to be involved in designing the systems, particularly consumers who really have the biggest stake of all. An example where teachers have had major input on a system is the system which Ruben talked about earlier in Texas, the statewide evaluation system. With that system, the state teachers association had a great deal of input. In fact, I believe they were almost entirely responsible for it with the result that, if you don't enhance that system to some extent, if you have a pulse, you're satisfactory. In other words, it is really geared in a way that it is very difficult to get an unsatisfactory rating on that system. In designing these systems, I think we have heard a lot of talking about issues of equity and fairness for teachers. We also need to keep foremost in our minds issues of equity and fairness for students. In the final analysis, most of us in this room, or at least a lot of us in this room, are evaluators, and really, one of our major responsibilities is to protect the interest of students or protect the interest of the clients of the public schools. Also, I was going to talk about what needs to be done before the center can do this; however, in the past two and one-half hours or two hours or whatever it was, you began to make progress, I think, on getting down some of the basic agreements that need to be sorted out before you can really begin to work in earnest on these materials.

One area of controversy which I promised Dan I would talk about, and I will, is the use of student performance data to evaluate teachers and administrators. The genesis of this in Dallas did not come from the education profession; it did not come from the evaluation department; it came from consumers. In other words, it came from the community. The community was absolutely outraged that there seemed to be no relationship between personnel evaluation and student performance on almost any variable you can discuss. I guess one thing that I would rapidly say is that when I talk about student performance, I am not only talking about standardized test performance, I am talking about things like dropout, things like graduation rate, things like attendance, things like the types of grade distributions that teachers give kids, the number of kids that teachers flunk, even down to such things as differential absences among teachers--that's a very informative statistic. Basically, you have seven periods in a day and sometimes when you look at the patterns over those seven periods you see that particular teachers have extremely high absence rates. In other words, kids just skip their classes. That is telling you something. You have got to go beyond the statistics and look at what that is telling you. That is telling you that, basically, kids are avoiding certain teachers' classes. Before, when in fact we got this major press from the community, which was then taken up by the board of education, of course the first thing that we did before attempting to get involved was to do a major review of literature. I have read the writing of many in this room. I would have to say that probably between 75 and 90 percent of the literature is opposed to using any type of performance indicators in evaluating personnel. There are a number of reasons that are given. I would also have to admit that, while I am philosophically in favor of this, in other words, I think philosophically it is indefensible not to talk about using student performance measures to evaluate teachers, methodologically I was extremely reluctant to get into the area because of the fact that it is just ripe with methodological problems. We did a review of the literature in terms of looking at what was going on across the country in the area, and basically in an attempt to convince the board of education that we really shouldn't do this, we documented a number of reasons why this is very difficult to do. I'm going to go down through some of those reasons just simply because I think it might be enlightening in terms of the direction that you seem to be headed.

One issue that often emerges is that it is difficult to measure long-term development of skills in terms of one-year periods. Obviously, the response to that is that you set up a longitudinal evaluation system. You don't look at one-year periods; you look at a longitudinal student progress-type situation.

Second, and the one that probably appears most often, is that really there is no instrumentation available for the assessment of diverse areas of achievement. What immediately comes up are things like music, art, physical education, etc. Our response to that has been you have to start somewhere. It may well be that you originally don't include music, art, physical education, whatever other kinds of things on which you don't have decent assessment.

A third issue, and one that really is a relatively thorny issue, is that the public schools, particularly at elementary school level, are not organized in a way that you can necessarily get at teacher effect. All teachers that a student is exposed to are really responsible for teaching and reinforcing reading. In other words, they are reinforcing reading through social studies, reinforcing reading through math, reinforcing reading through science, etc. Our solution to that at this point, at least our tentative solution, is that we may not have data on specific teachers. The lowest we can go may be pods of teachers or groups of teachers, in other words essentially looking at a number of teachers who have major impact on students. This is particularly a problem, as I said, at the elementary level; at the secondary level it is not so great a problem.

Fourth, there is a great deal of ranting against norm-referenced standardized achievement tests. I would emphasize here that when we talk about performance measures or measures of outcome we are not only talking about norm-referenced achievement tests. We are talking about criterion-referenced tests; we are talking about, in some cases, performance tests; we are talking about other variables like attendance, like grades, like percent passing, like percent of students in honors programs. We are even, in terms of our effectiveness indices, looking at graduate follow-ups, at what students do when they graduate from high school, and beginning to look at the effectiveness of our programs in those areas; so that when I'm talking about student outcomes, I am talking about a very broad array of outcomes. In one of the papers I did for AERA, we listed about 40 outcomes in terms of the kinds of things we are attempting to consider. In addition to that, there is the idea that standardized achievement tests do not reflect the full range of instructional goals. We have responded to that by having a series of tests, 143 final course examinations called Assessments of Course Performance, which are given in grades 7 through 12 in everything from reading improvement to calculus. These are used as part of the overall system in terms of looking at (a) the impact of the school and (b) the impact of the individual teacher. One other thing that was able to persuade me to really start getting into this was that you see a lot of literature about what is wrong with trying to do this, but you don't see a whole lot of literature about what is wrong with not doing it. Speaking from my own personal perspective, I have a daughter who is currently in medical school and is still suffering from the effects of chemistry according to Ms. Bowen in the eleventh grade. In other words, there have to be some basic standards. There have to be some basic standards that you are going to hold teachers accountable for teaching and you are going to hold students accountable for learning. And the thing that really is amazing when you start looking at our ACP patterns is that where the kids are really suffering is not so much in reading improvement and the lower-level courses. Where the kids are really suffering are courses like calculus, biology, physics. That is where we seem to have the greatest mismatch between what is expected to be taught and our teachers' ability to deliver that instruction, at least across the district.

The fifth issue that we needed to address is the issue that what the student brings to the classroom in terms of ability, home and peer influence, motivation, and other types of background influences is very powerful in affecting academic achievement. This is what Ruben and I were talking about a minute ago. In other words, in order to have a really fair system, you have to look at what the student is like when he comes in, look at what the background factors are, the factors that affect achievement, and then try to factor those out in the overall equations. We have been relatively successful in doing that at the school level. We have not been very successful yet in doing that at the teacher level, and I'll talk about that in a minute.

Another point that keeps emerging against using student outcome variables is, first, that the statistical methods used to control for nonteacher factors can't take into account all relevant factors, and secondly, you can't explain them to anybody. Well, one cannot take into effect all relevant factors that may be true, but they can surely take into effect a lot of factors and they can surely do a better job at attempting to provide some credible data for evaluation than if you do not try to take those factors into account. Also, in terms of explaining them, our experience has been that it has been very easy to explain the methods. You don't necessarily have to explain the statistics and the mathematics. What you explain is what the method does, in other words, what the effects are. In fact, we are to the point now in our feedback to schools and to teachers, that rather than giving them actual scores, we supply skills analyses and what have you. We actually provide them with standard scores on everything broken down by ethnicity, by gender, by economic status, limited English proficiency status, and all the interactions between those; with scores that have a mean of 50 and a standard deviation of 10 so that they can look quickly and see who they aren't serving. In other words, any scores below 50 indicate that they are obviously failing to meet the expectations in terms of who they are serving. This has proved to be a very powerful tool.

Another issue that has come up and the one that we are still dealing with (or still trying to deal with) is a very thorny problem; it is the degrees of freedom issue at the teacher level. Because of that issue, essentially you don't have large enough n's to have stable estimates on what the predictions ought to be, etc. The approach we are taking is that we have convinced the Board that we do not have to have little teacher effectiveness indices on every teacher. The principal, under our site-based management plan, is the chief executive officer at the school. We are providing the principal and teacher, the individual teacher, with a great deal of information in terms of what their students look like when they come in, what they look like when they leave, and what some of the relatively unique and questionable patterns such as absences, failures and what have you are. While that does not directly relate to teacher evaluation, it is obviously being used in teacher evaluation. The next step that we are now being pushed into is to take the Domain 13 on the Texas Teacher Appraisal System and apply student performance measures to that domain, so that there is an actual one-to-one link between performance and evaluation.

One of the major problems that you have in attempting to get involved in these systems, and one of the problems that you are going to have to face, is which variables to use, how to measure the variables, how important they are, etc. I can again only talk from the standpoint of a local school district; our consumers are the ones who determined what the important variables were, and our consumers are the ones who are determining the appropriate levels of those variables. Now that is in conjunction with teachers, administrators, community members, and even the teacher unions. We don't really call them teacher unions, but the teacher organizations are all involved in that. But the consumers have the heavy role, and the reason is that the schools basically belong to the consumers, to the people. They are the ones who pay the taxes; they are the ones who pay our salaries; they are the ones who pay the bills; therefore, what they believe is important. We need to scrutinize, and obviously one of the things that they feel is important, is standardized achievement tests. It would be very difficult to establish, to develop, a system like this that did not include, as one component, standardized achievement tests, simply because of the fact that the public puts so much stock in them.

You need also to look at ways to interrelate all this, such as where to start, etc. One way to do this, and I think one way that this project might benefit through progressing, is to start at the school level. In other words, you start looking at effective schools, or defining effective schools at the school level. You work down; the principal evaluation is based in part on how effective that school is. In fact, in Dallas it is based a lot upon how effective that school is, and then the teacher evaluations feed into that, and it is all part of a general system. All are looking at the same outcome variables. I think you mentioned that yesterday, Michael; basically you don't have differential outcome variables for different levels. You all are attempting to accomplish the same thing. The principal is accountable for the operation of the school, and, as such, he or she is also accountable for making determinations about teachers based on the data that they have.

How can CREATE help this entire process; what can CREATE do to help; or, as someone stated yesterday, why should a school administrator care to look at the materials? I think school administrators, teachers, and educators in general will care if the products can make their jobs easier. I think everybody wants to do a good job. I've run into very few people in my travels who really did not want to do a good job. The problem is the amount of effort, above an already overextended situation, that can be put into doing a good job. That is why I talked originally about the materials that need to come out of this process being simple and straightforward. They need to be very simple, very straightforward, very easy to implement, very easy to understand, if possible, and I realize that this is asking a lot, but, if possible, not providing many alternatives where they have to read lengthy background data and make decisions. Essentially, we try to look at what's the best practice or what is the best way to approach this.

In terms of the evaluation/research community, again what I think CREATE can produce, and what, from looking at the material in Arlen's Table 7, CREATE is looking at is the research behind the products produced for the practitioners. I think Michael yesterday said he had nearly 3,500 sources in one of his reviews of the literature. That's very interesting and informative for the research and evaluation community, but not at all informative for practitioners. They would really not be interested in that. Another thing that CREATE can produce is a scholarly paper on the relationship between personnel evaluation, product evaluation, school evaluation, etc., trying to tie the whole thing together. I don't think there is much point in producing that for practitioners, because I think the people that are going to do that are people from the evaluation/research community. And, finally, I guess after listening to the dialogue this morning, I think it would be very useful to have a scholarly paper also produced comparing and contrasting the various personnel evaluation models so that we can get a really good feel for where the similarities are and where the differences are.

For the policymakers, I think that what needs to be produced is an inventory of the systems that are coming out of CREATE under the service to practitioners and service to the evaluation/research community as well as ways they relate to accountability. Again, I think the major thrust is ways that all of this relates to making educators accountable and making schools accountable. Outside of accountability, I think that's about what I have to say. Thank you.

RESPONSES TO WILLIAM WEBSTER'S SEMINAR SUMMARY

The following are responses from Dr. Daniel L. Stufflebeam and other seminar participants to Dr. William Webster's summary of the seminar.

Stufflebeam: "Thank you very much. I was wondering if you have any further ideas about what we should do for the policy group. You gave a rich set of ideas for the practitioner, and I know that there is considerable interest in this group about what we can and should be doing to help, for instance Washington, in developing an agenda for research and development in this area, and what we can and should be doing for state departments and school districts to inform them, to get them to buy in, and your brief commentary doesn't reflect your very deep experience in this area, but I would like to push you to say some more about it."

Webster: "My experience in this area is actually rather limited. You know I have a great deal of experience in the area working with a local board of education. What I think has led to some degree of success in getting support for evaluation activities is to make what it is you are attempting to do real. In other words, explain to them ways that this can help them better monitor what's happening in the district. I have very limited experience with working at the state level. In fact, right now I guess we are engaged in these kinds of discussions with the Commissioner of Education in terms of why we need to adjust outcomes as part of a system. I suppose one thing that this reminds me of and I need to say is that there are really two parallel systems. One is the one in which we are talking about adjusting outcomes. The other deals with absolute goals based upon where students are. We can't ignore one in favor of the other. When we talk about school effectiveness indices, teacher effectiveness indices, or whatever level you happen to be talking about, the way we are implementing it, that's a normative concept, so that being last in a year when the district has a good year, achievement wise, is not as bad as being last in a year when the district has a terrible year. So, again it is a relative kind of thing. I would say, if you are asking for my experience working with policymakers, I would say number one is to explain succinctly the implications of what you are proposing to do. The methodology is not as important as the conceptual framework when working with policymakers."

Stufflebeam: "Including your local board."

Webster: "Yes. I would say primarily to keep it short and to the point. As I rather facetiously said at the beginning of this presentation, anything that goes over ten minutes loses the audience. They really want to know. Yes, I know it sometimes is over ten minutes; I don't mind when you look at your watch; it is when you look at it and start hitting it that it bothers me. Mainly I would say to keep it short, keep it to the point, keep it accountability-oriented. Unfortunately, policymakers really don't care a lot about improvement. What they care about is accountability. It is my own personal belief that it is extremely counterproductive to have an accountability system without a built-in method for improvement, and this is, again, one of the main things that we are talking about with the state. The state currently has a testing program, the Norm-referenced Assessment Program for Texas, which does not provide skills analyses, so that essentially you get accountability information but you don't get any type of information to help you improve. To me that is wrong. If you are going to hold people accountable for something, you also need to provide them with the method to improve or provide them with the knowledge of what it is that they need to improve."

Audience: "You've had substantial experience with the concept of merit pay and merit school . . . I wonder if you could provide some information about that. What is your assessment on the liability of doing merit-school evaluation as opposed to individual merit evaluation?"

Webster: "I think individual merit pay evaluation in many ways is counterproductive. I have fought very hard in Dallas to keep it at the school level because, essentially, when a school is an effective school, everybody benefits. This, of course, is in concert with providing the principal with the necessary information to make decisions within that context in terms of who is contributing and who is not. There is a major advantage, I think that we've experienced (should I talk a little bit about the outstanding school performance awards?) One of the things that the community suggested that we do is not only to figure out a way to define effective schools, or outstanding schools, but also to reward the staffs in those schools. So, last year in the top 20 percent of schools in the district, their teachers received a $1,000 bonus, and their support staff received a $500 bonus."

Audience: "Was there anything for the principal?"

Webster: "The principals got $1,000. The entire professional staff did. We have felt that there have been many positive fallouts of this program. One thing that we have been able to notice that has been demonstratively different is that there seems to be a great deal more help within buildings. In other words, teachers are more willing to help each other. If I'm a social studies teacher, and the one next door is floundering, it seems like teachers have been more willing to help, to work with those people within the building. Another interesting spin-off of it, at least in the elementary schools, is that principals seem to organize away from their poor teachers. If, in fact, there is a major weight on reading achievement, then generally the best teachers are the ones that are really responsible for teaching that in the schools. Another major component of this program, and probably the most successful part of the program is, if you ignore the methodology and whatever, the accountability task force. This is a permanent task force appointed by the superintendent that consists of all the major actors in the community that make all the decisions relative to (a) what is important; (b) how to weight what is important (in other words, what are the differential weights); (c) what types of methodology do you use? Obviously, in the early meetings of this task force, some of the meetings went six or seven hours, while we talked about the relative merits of this form of regression analysis versus step-wise regression analysis versus canonical correlation versus hierarchical linear modeling, in terms of actually showing the kinds of results from these various models. Probably the most interesting discussions that I've heard since I've been in Dallas (and that has been 23 years) in terms of what education is all about, have come out of that Accountability Task Force. I mean in terms of what it is we are really attempting to accomplish, with the parents debating with the professionals, or debating with the business community, back and forth. It has been a very positive process."

Audience: "Can you say something about what it takes to be in the top 20 percent criteria?"

Webster: "Effectively, the criteria differs at different levels. The competition, if you will, is run K-3, K-6, 7-8, and 9-12. We don't run any across levels because you have different student-growth periods across levels. What we effectively do is run a series of prediction models on an individual student basis. This system encompasses two and one-half million equations on an individual student basis where we look at prediction, based upon individual student histories, and upon where they ought to be on particular variables. Then, we look where they actually are and apply that back to the schools. To oversimplify it, the schools that have the most kids that exceed expectations on the most variables, given the weighting in the system, will be the ones that win. They are the schools that take what they have and do the most with it."

Audience: "It's a question of variables?"

Webster: "It's value added."

Audience: "It's the variables that I am concerned about."

Webster: "Okay, the variables. I would test the basic skills or the norm-referenced assessment program in reading, language, and math, with assessment, of course, on performance grades 7 through 12, 143 different courses, plus graduation rate, grades 9 through 12, which is the percentage of ninth graders who graduate within five years. We examine promotion rate in the earlier grades, elementary schools and middle schools, which is essentially a percentage of kids promoted. We check student attendance; again, student attendance is not student attendance against the district average, it is student attendance against what the history has been, with the idea that you are attempting to improve individual student attendance. Enrollment in higher-level courses, honors, advanced placement, etc., SAT scores, and percentage of students taking the SAT are also examined. Beginning next year, graduate follow up, or what the students do when they graduate from school will be added. We didn't have dropout rate in this year's study because there has to be a time lag, so schools have to be told ahead of time. You don't really get dropout rate until November, and the results are announced August 15th . . . a lot of variables. In addition to that, we are trying to figure out the ACPs; we have 21 ACPs next year that will have performance tests on them. What we are essentially doing is trying to figure out the scoring protocols. We are obviously going to have to let the schools score them, because we would get into a tremendous amount of resources if we tried to do it. But then we will conduct a sampling basis to determine their reliability , and then weight the performance part of the ACPs by that reliability in the effective new system."

Audience: "Bill, how would you account, in your school district, for the fact that you have a higher percentage of kids not going on to college, maybe 60 percent going on to college and the other 40 percent not. How do you make sure your weighting system is not tilted toward those schools with the kids going to college?"

Webster: "Everything is really on an individual student basis, in other words, what we predict on an individual student basis. It concerns not only going to college, but also unemployment, military service, and a whole range of options."

Audience: "What degree of repeating do you find in these kinds of general classifications?"

Webster: "That is a different question than I thought you were asking. Usually, it is pretty consistent. What happened, and something that was an unanticipated outcome was that our school board decided that they wanted to talk to the lowest 20 schools in terms of the principals. So we had a retreat on Friday and Saturday for 2 different weeks. Each school had about an hour and a half in which they basically discussed what they were attempting to do to remedy their situation. What really is amazing was the outcome of that; we did not produce the actual effectiveness indices, but what we did produce was cohort patterns on the various achievement tests. A lot of the principals were saying originally that they lost because of the fact that they had too many white kids, and white kids were differentially weighted against. Well, of course that was not true. They lost because their kids did not maintain their growth expectations. But, at any rate, almost all of them had a reason; for example, we had one extreme example which was Frazier Elementary School, a K-3 school. Grades 1 and 3 looked absolutely terrific, in other words exceeding prediction. Grade 2, from the pretest to the posttest, the cohorts of students lost an average of about 17 NCEs in reading, language, and math, which of course is why they ended up ranking low, because that was a tremendous loss. Actually, the principal was able to explain what happened. Essentially, they didn't have any regular teachers in Grade 2; they were all substitutes, and they had three or four different substitutes during the year, etc. To me, to hear these reasons time after time, about what actually happened, really added in my mind to the validity of the process. The other question that I thought you were going to ask . . . when we did the same type of thing in 1984, (only at that point and time we were much more limited), we were just talking about standardized test results, because that is what the community was really interested in. We did the same thing, monetary awards, the first year; the second and third year we no longer had the monetary awards. The original system was designed because we had a $3 million windfall which we wanted to put into personnel, but we could not maintain it. So we had to do it on a one-time-only basis. Back then, based on those models, which are similar to the ones we are using now, it was very consistent; in other words, schools tended to stay up there. The main thing that impacted schools drastically was changing the principal. This was quite interesting. That was something that we did not predict, because the way in which the original methodology was done, we were actually using 2 or 3 years of student history on each variable so that individual student growth curves were being looked at. The winning schools won by accelerating their student growth curves, so, rationally, the next year would be hard to meet, but, in fact, it was not; it was very consistent. Schools tended to stay high. The second year, about 70 percent of the same schools were at the top."

Audience: "Is that a particular grade level? Or is it across grade levels?"

Webster: "No, I'm talking within K-3, K-6, 7-8, and 9-12 grade configurations."

Audience: "Do you test all the grades within each of those blocks?"

Webster: "Yes. If you run methods like this across grade levels, your high schools are always going to lose, because they do not have the same opportunity to accelerate growth curves that your elementary schools do."

Audience: "I just want to comment. You mentioned that the individual merit-based systems were not necessarily positive. As the developer of a major individual merit-based system, not by choice by the way, my own personal experience, and by reading the literature, is that the best possible individual merit-based system will have neutral effect. Most systems are not the best possible, most systems have a negative effect as their net impact."

Webster: "Yes, and I think one of the reasons I made that comment was because in education what we are attempting to do is get some collegial relationships within the school, that is, teachers helping teachers to do a better job. One of the ways to do that is not to run Teacher A against Teacher B, in terms of competition."

Audience:	"Have you ever thought about the possibility of doing criterion-referenced merit school awards? If you can find a certain level of growth curves you are after. Then basically every school in the district would have the chance of achieving that and being rewarded."
Webster:	"Yes, we thought of that. Our problem is that we have had some difficulty in setting realistic expectations. We have been talking about using multilevel modeling to do that. In other words, with multilevel modeling your cross-district equations are one level, but you have within-school equations in the other level that really look at the school against itself, and whether or not it can, in fact, meet or exceed its expectations. We are thinking about that. The other thing that we are currently thinking about is that the current system is based on continuously enrolled students, mainly because the educators absolutely insisted on that: students that were exposed to the educational program. Unfortunately, in Dallas, the district as a whole has a 32 percent mobility rate. That is not as bad as it sounds; the 32 percent is contained in only about 18 percent of the students. Now we are looking at equations to try and bring those kids into it; in other words, looking at the system-wide relationship between attendance and achievement to see if we can plug that gap, because, right now, obviously, we have a hole there."
Audience:	"What has happened in Dallas is an example. I think Dallas is probably one of the most advanced systems within the state in getting an outcome based on the evaluation model from schools and school districts. I think what has facilitated that has been largely due, besides the local leadership of the superintendent and some of the school board members, to the framework that has been established by law. All of the indicators and variables that he mentioned for measuring the school performance are required to be used by the state. Your norm-referenced test, your criterion-referenced test, your graduation rates, and every one of those variables is an expectation as far as the criteria we are now using. Some recent conclusions now, as a result of having addressed the problems of the school finance there in Texas, the issue of equity, it seems like in the last six years or so, every time that comes around and goes to the court and comes back, that accountability packages are added. We just passed some legislation this past week, in fact Monday or Tuesday, I believe, it was signed by the governor, which enhances a lot of what is already very important in terms of the academic excellence indicators of whole school districts. Now, very specifically, the fliers require that school districts evaluate the performance of the superintendent and the performance of the principal on the basis of those academic excellence indicators, which are basically the variables that Bill has outlined. It also very exclusively outlined a very strong set of consequences or sanctions that the commissioner must oblige, creating more of an expectation now by law, rather than by a choice by the commissioner. The commissioner must establish standards of clearly acceptable and unacceptable performance and then have the ratings imbedded in that. For those school districts that are clearly unacceptable, and campuses that are clearly unacceptable, an automatic lowering of their accreditation status occurs. Following that lowering of the accreditation status, there is a very comprehensive review by an intervention team composed of peers. The intervention team, both at the district and the campus levels, would have the purpose of determining the extent to which the local campus or district would reach a clear understanding of the problem, and a clear designation of how they are going to correct the problem. This gets into a whole bunch of criteria within the planning process. The sanctions that can be applied, or other intervention measures that can be applied at the determination of that intervention team, both in the district and at the campus levels, are not having to do with the accreditation status, the accreditation status must remain lower until

progress in student performance is demonstrated in the next round of the measurement of these variables on a yearly basis. Such sanctions and interventions as removal of principals, and entire faculty, starting from scratch, at a campus level; removal of campus from under the jurisdiction of the school board by designating them more the managers, appointed by the commissioner within the community; removal of school board members and taking over school districts if they operate poorly in school systems; or applying the management teams monitoring systems; or actually annexing or consolidating school districts after a designated period of time, again is in the law. Failure to demonstrate the expected growth toward those standards is shown by those schools. It is an extremely strong outcomes-based or results-oriented accountability system that is having tremendous implications toward how we rethink the whole evaluation concept, both at the school level and at the education personnel level."

Webster: "The major debate we are having and what this system is all about, is attempting to distinguish the difference between the student who is born on third base and the one who hits the triple, in terms of trying to look at what the schools do with what they get."

Audience: "In terms of that question, what 20 percent of the school will end up in your top? Are you attempting to synthesize anything they have had to do in all of the successful schools in terms of characteristics of them, so that they might provide information to schools that are not that successful?"

Webster: "No, I told Curriculum and Instruction to do it."

Audience: "Is somebody doing it?"

Webster: "Yes, they are trying. The interesting fallout of this is that we get schools from all parts of town with all kinds of demographic characteristics. The number one high school in this district, in terms of effectiveness, is Lincoln High School, which is 100 percent African-American and probably 40 percent low income, etc. You can walk into Lincoln and then go and walk into other schools that serve similar students; there is no comparison; there is absolutely no comparison. The system has a lot of validity which is absolutely necessary. In other words, if we were to come up with this formula which produced schools that were way out, we would probably be in a little bit of trouble."

Audience: "How long is the curriculum team going to attempt to decide what it is that caused the difference?"

Webster: "Presumably, 22 years. The latest round is probably the last couple of months. We know what caused the difference in a lot of cases, because in almost every case you get a very strong principal. In the case of Lincoln, you have a strong principal. He is not necessarily a strong instructional leader, but he creates an atmosphere in which teachers can teach and he supports his teachers."

Audience: "If that is true, and obviously it is, in our place like Maryland, we are attempting to use some of those pieces of knowledge in our developmental academies in terms of what do people who are the future principals or practicing principals do when you present them with some of these situations?"

Webster: "Yes. Obviously, it has a lot of implications for a lot of things other than accountability. One other part of it, which I forgot to mention but which is very important, is that another part of this (and one that we have gotten into unwillingly and really did not plan on) is that we are now into a lot of staff development. We, the Research Department, are being requested by the schools to give them help in things like interpreting data and how to modify instruction based on data. Another area in which we are trying to branch out is to broaden the base of indicators. We continually try to broaden the indicators. We are teaching many of our teachers how to do protocol analysis, so that, essentially, when in the evaluation process they get results, they have some other results to either debate the outcomes or to defend their indices. We are teaching our teachers to performance-test. My own personal opinion, again based on a lot more years than I care to remember in the public schools, is that the best way to get something taught is to try to measure it. In other words, if you do not measure it; it tends not to be taught. I've seen this often looked at as one of the major limitations of standardized tests, because if you use standardized tests, obviously that is what teachers try to teach. But if you broaden the measures enough, and if you look at alternative ways to measure and alternative ways to look at progress, then, in fact, you get a payback in terms of the way teachers are delivering instruction."

Audience: "The general view was that everybody thought you could not really do this kind of outcomes-based evaluation. I feel exactly the same about merit pay. I think you gave up too early on merit pay. I have been around a good many merit systems since the college level, for teaching only, nothing to do with research, and they worked extremely well. We got people from across the whole spectrum, and everybody said you will never get anybody from math or physics. We have had absolute response to that subject matter. So I do not think we want to give up that. There were only a few people, and I don't want to single anyone out. That is so implausible since we are living in the university where on the basis of research you are being singled out and promoted all over the place."

Webster: "Yes, but we are talking about completely different environments. In our environment, we are talking in many cases about a lack of skills where essentially the more aid we can get at the point of instruction for teachers that are lacking in specific skills, the better off we are going to be and the better off the institution is going to be."

Stufflebeam: "One of the mandates that we got from Washington last year was, rather than continuing to engage John Sandberg in going across the country and interacting with ten states a year on our behalf, we were asked to select four states in this country to interact with as much as we could. We selected Texas as one of the four; Texas did not select us. I should add we just selected them, one of the things we tried to do in this institute is sort of feature Texas in that respect. I am really pleased that Ruben came and of course Bruce is always with us. We are delighted that Bill came and gave such a nice summary and a very useful perspective of the work that is going on in Dallas. I think that there is real potential for CREATE in continuing to study and interact with the Texas Education Agency and the Dallas public schools and hope that this can be the beginning of some networking that will occur between our various projects and with the good work going on in Texas. You heard, Bill, that there is more than one person in this group who is interested in having you publish your papers in some of the journals that we are using as our outlets."

SEMINAR EVALUATION REPORT

Stanley Nyirenda
Western Michigan University

This report is not a record of the seminar. Rather, it is an analysis of the perceptions of the participants about how projects undertaken by the Center for Research on Educational Accountability and Teacher Evaluation (CREATE) may be integrated in a coherent way. It is a report of what participants have identified as helpful strategies for CREATE to use in developing a cross-cutting theory that draws together the common elements in all its projects, as well as in developing products that are useful to practitioners and supported by policymakers. The report is intended to supplement the record of the seminar and to provide CREATE's director and associate director with some suggestions for planning the tasks to be accomplished by the Cross-cutting Theory Project over the next two years of the program.

The report is based on two evaluation instruments. The first instrument, administered to participants at the end of the first day, elicited the seminar participants' perceptions of the presentations and group discussions about what CREATE should do in the Cross-cutting Theory Project. Apart from getting feedback from participants on how they viewed the quality of presentations and group discussions, this instrument was intended to provide information that would help the seminar chair and CREATE's director revise the agenda for Day Two, in light of the experiences of Day One. The second instrument, administered at the end of the seminar, solicited participants' perceptions and comments for a summative evaluation of the seminar. Both instruments had selected response items using 4- or 5-point Likert scale type items and open-ended questions. The perceptions of participants regarding the first day's agenda are presented before those of Day Two.

PERCEPTIONS OF SEMINAR PROCEEDINGS - DAY ONE

The results of participants' responses are discussed under presentations, planning sessions, impact comments, and recommendations by participants for both presentations and planning sessions. Differential semantic scales are used to describe participants' feelings about presentations, their participation in the planning sessions, and how helpful they believed such discussions to be in moving CREATE toward its goal of developing a cross-cutting evaluation theory that links together its research and development activities in a way that maximizes the improvement of the practice of evaluation of school personnel and systems. The responses are presented in Tables 1-5 for the closed-ended questionnaire items and described for the open-ended questions. It should be noted at the outset that not all 27 participants completed the evaluation survey. The highest number of responses for any item was 21, and the lowest for the open-ended items was 0.

Presentations. The responses for each scale point in all the tables are expressed in both frequencies and percentages to facilitate comparison. In Table 1, the results of the participants' ratings of the presentations, on the whole, show high ratings. Except for "Politics of Teacher

Evaluation," which was rated poor by 15 percent and fair by 45 percent of the responding participants, the other presentations were rated good to very good by between 85 and 94 percent. The overview of seminar objectives was considered good to very good by 85 percent, the overview and analyses of approaches and findings from CREATE's projects by 90 percent, and "Foundations of Educational Accountability" by 94 percent. The responses relating to CREATE's projects suggest that it is clear about its objectives, especially at the individual project level, and also that it is clear about the need for a cross-cutting theory that would tie the different projects together into a coherent program.

Table 1: Participants' Perceptions of Presentations

	Frequencies in Percentages					
Presentations	n	Very Good	Good	Fair	Poor	Very Poor
Overview of seminar objectives	20	45.0 (9)	40.0 (8)	15.0 (3)	-	-
Overview and analysis of approaches and findings from CREATE's projects	20	40.0 (8)	50.0 (10)	5.0 (1)	5.0 (1)	-
Foundations of Educational Accountability	17	35.3 (6)	58.8 (10)	-	5.9 (1)	-
Politics of Teacher Evaluation	20	10.0 (2)	30.0 (6)	45.0 (9)	15.0 (3)	-

Key = Numbers in brackets represent the actual frequencies

Planning Sessions. Participants were divided into four groups to discuss a possible work plan for the Cross-cutting Theory Project. At the end of the day participants were asked to express their feelings about the group's discussions concerning the work plan. The specific agenda suggested by CREATE's director included discussions on (1) relative priorities to be assigned to serving different constituent groups, such as evaluation researchers, school practitioners, and teacher educators; (2) what outcomes the project should seek, e.g., issue papers, materials for training teachers, school practitioners, and teacher educators; (3) relative project emphasis on various project activities, e.g., cross-cutting issue seminars, product development efforts, a combination, or any other activities; and (4) relative priorities to be assigned to cross-cutting issues, e.g., politics of evaluation, use of student performance measures to evaluate teachers, and/or achieving sound teacher evaluation in the context of collective bargaining agreements. Participants' responses are shown in Table 2.

Responses regarding participants' perceptions about the discussions on the four planning issues previously identified by CREATE's director showed that no participant felt that such discussions were very poor; 5.2 percent and 5.6 per cent considered the discussions on the "relative priorities to be assigned to serving different constituent groups," and "relative priorities to be assigned to cross-cutting issues," respectively, to be poor; between 66.6 percent and 74.8 percent felt that their group discussions on each of the four planning issues were either good or very good; and between 21.1 percent and 27.8 percent felt they were fair (see Table 1). In Table 2, about 33.3 percent felt that discussions about relative priorities to be assigned to cross-cutting issues were either poor or fair. This suggests that the least agreement was on which issues should receive most attention. Participants seemed to have similar perceptions about the other planning issues as suggested by the narrow range in favorable responses (i.e., responses of good or very good group discussions accounted for 73.7 and 74.8 percent).

Table 2: Participants' Perceptions of Planning Discussions

Discussion Topics	Frequencies in Percentages					
	n	Very Good	Good	Fair	Poor	Very Poor
Relative priorities for constituent groups	19	10.5 (2)	63.2 (12)	21.1 (4)	5.2 (1)	-
Outcomes to be sought	18	22.2 (4)	55.6 (10)	22.2 (4)	-	-
Relative priorities to be assigned to cross-cutting issues	18	11.1 (2)	55.5 (10)	27.8 (5)	5.6 (1)	-
Project emphasis on various activities	19	21.1 (4)	52.6 (10)	26.3 (5)	-	-

Key = Numbers in brackets represent the actual frequencies

As shown in Table 3, 89.5 percent of the participants responding found both the presentations and planning discussions to be stimulating or very stimulating. Only 10.5 percent were uncertain. At the same time, 33.3 percent were unsure about the helpfulness of both the presentations and planning discussions with respect to furthering CREATE's evaluation theory development effort.

Table 3: Participants' Overall Ratings of the Day One Proceedings

		No. of Responses	Percentage Response
Participants' overall ratings of Day One's presentations and planning sessions	Very stimulating	5	26.3
	Stimulating	12	63.2
	Somewhat stimulating	2	10.5
	Not stimulating		
	Very unstimulating		
	Total	19	100%
Helpfulness of the presentations and planning sessions to CREATE's effort in theory development	Very helpful	1	4.8
	Helpful	13	61.9
	Somewhat helpful	7	33.3
	Not helpful		
	Very unhelpful		
	Total	21	100%

Suggestions for Improvement of Format. Only 2 of 21 responding participants made suggestions as to how the format might be changed to improve the following day's discussions. The two responses were "Need to respond/discuss more concrete/definitive issues as they relate to the future of CREATE" and "establish concrete expectations for outcomes." The majority of respondents preferred continued exploration of ideas on what they perceived to be the most beneficial way for CREATE to tie its projects together in a coherent program that is supported by both theory and practice. Although no suggestions were made about the agenda for the second day of the seminar, participants in each of the four planning teams had agreed to continue discussing issues identified from their first day's discussions.

OVERALL SEMINAR EVALUATION - DAY TWO

The evaluation instrument for Day Two represents participants' perceptions of overall seminar proceedings and the logistic support provided, as well as feedback on weaknesses and strengths and suggestions for future activities of the Cross-cutting Theory Project. Tables 4 and 5 summarize the responses for the closed-ended items, while the open-ended items are described briefly.

Table 4: Participants' Overall Perceptions of Conference Proceedings

		No. of Responses	Percentage Responses
The extent to which participants benefited from the presentations and discussion groups during the seminar	Very great extent	9	45.0
	Great extent	10	50.0
	Some extent	1	5.0
	Total	20	100%
Participants' overall assessment of the success of the seminar in identifying, and delineating issues, and developing a work plan for a cross-cutting evaluation theory for CREATE's projects	Very successful	4	19.0
	Successful	9	43.0
	Somewhat successful	8	38.0
	Unsuccessful		
	Very unsuccessful		
	Total	21	100%

Most participants felt they personally benefited to a great or very great extent through their participation in the seminar (95 percent). Overall, 62 percent viewed the seminar as very successful or successful, while 38 percent were not certain and rated it as fair.

Participants were most pleased with the meals provided during the seminar and the facilities for activities other than the seminar (100 percent rated them as good or very good). Discussion papers and presentations came next (95 percent), followed by equipment as applicable (93 percent) and planning team sessions (86 percent). It must be noted that very little use was made of equipment such as overhead projectors, video, and flip charts as confirmed by the low responses to this item (15 responses). However laptop computers were provided at each group table for use by discussion summarizers.

Table 5: Overall Rating of Materials, Facilities, and Equipment Used at the Conference

	n	Very Good	Good	Fair	Poor	Very Poor
Discussion papers/presentations	21	57.0 (12)	38.0 (8)	5.0 (1)		
Planning team sessions	21	48.0 (10)	38.0 (8)	14.0 (3)		
Facilities for activities other than the seminar, e.g., recreation, lodging	20	65.0 (13)	35.0 (7)			
Meals during seminar	21	57.0 (12)	43.0 (9)			
Equipment as applicable	15	40.0 (6)	53.0 (8)	7.0 (1)		

Key = Numbers in brackets represent the actual frequencies

Open-ended Item Responses. Not all participants responded to the three items of the second instrument that solicited feedback on what was most productive at the seminar, the weak points, and how CREATE could use the ideas and materials produced during the seminar to improve the organization of future seminars. A transcription of the suggestions, comments, and recommendations made by participants is available in a separate document. Only those comments and suggestions made by the majority of participants are provided in this report.

> *Most Productive Aspects of the Seminar*. First, a significant number of the participants who responded to this item considered the professional exchange of ideas among individuals with often divergent perspectives and viewpoints in the field of evaluation, and the resulting networking, to be most productive. Second, presentation, discussion, and synthesis of a select number of personnel evaluation models (i.e., Professional Support Personnel Evaluation Model - Stronge and Helm; Teacher Development Model - Iwanicki; Praxis Theories: Professional Assessment for Beginning Teachers - Dwyer; and Duties-Based Model - Scriven) and especially the identification of the similarities and differences were considered to be an important step toward developing a generalizable model for personnel evaluation. Third, papers and presentations identified fundamental issues that must be considered in the theory development effort of the cross-cutting project. Finally, small group discussions provided members the opportunity to take ownership of ideas and relate them to their own experiences; this was considered a strong outcome of the seminar.

Weak Points. Some participants identified the following as weaknesses:

1. information on the current status of projects/models for personnel evaluation being presented too late to guide discussion
2. too much emphasis on specific products at the beginning of the meeting, use of committees to develop plans
3. group members had their own agendas for the cross-cutting theory project which were not synthesized
4. too much time spent getting separate ideas from the groups

Utility of the Ideas and Materials. Several suggestions were made to improve the organization of future cross-cutting theory project activities:

1. summarizing materials and making them a part of CREATE's documented history; using such materials as a basis for developing future "cross-cutting" seminars
2. putting more emphasis on exchange of ideas from the work of CREATE including models that have generalizable impact, rather than directors' presentations of their progress
3. providing sufficient time for careful presentation of models and papers for brief discussion and written feedback to presenters
4. breaking into group discussion early on the first day and reducing the duration of the entire seminar from three to two days, reviewing group suggestions while paying attention especially to conflict and using suggestions made through group reports
5. offering a series of three seminars/year to guide development of cross-cutting products (kit and sourcebook) involving at least the active members of the seminar and starting such seminars with a prepared position paper that addresses key issues

GENERAL DISCUSSION AND RECOMMENDATIONS

The overview of CREATE's projects explained how they are perceived to relate to one another by the center, their achievements to date, and how adjustments have been made to reflect the changed priorities of the program and the funding agency, U.S. Office of Educational Research and Improvement (OERI). For example, there were four programs in the original configuration: improvement of teacher performance evaluations; improvement of evaluations of administrators, support personnel and schools; products and services/dissemination; and theory development and special projects. Currently, one of the original programs has split into administrator and support personnel evaluations and improvement of school evaluations to ensure that each of these important aspects of evaluation receive adequate attention. In addition, the project years for each program presented in the overview demonstrated the weighting that is assigned to each program over the 5-year funding period. While teacher evaluation has the highest number of project years (12), cross-cutting theory and special projects comes second with 9 project years and total funding of $200,000 over a two-and-a-half year period. Participants found the background information useful and important for orienting their discussion. Any future cross-cutting activities will benefit from a brief overview of CREATE's work to focus discussions on developing a theory of evaluation that unites CREATE's projects in a sensible way.

The two discussion papers on "The Foundations of Educational Accountability" (Scriven) and "Politics of Teacher Evaluation" (Glass) and the discussions of them clarified the underlying concepts that link personnel to school evaluations. Evaluation of schools implies evaluation of personnel who work in those schools. Educational accountability and the politics involved in teacher evaluations that are undertaken by administrators who may lack both knowledge of effective teacher and school evaluations as well knowledge of effective teaching in disciplines unfamiliar to them, may provide common concepts necessary for building an effective generalized model. Such models can be used to evaluate educational personnel at all levels (administrators, teachers, professional staff, as well as university personnel) and to link such evaluations to student outcomes. The four models (Praxis by Dwyer, Professional Support Personnel by Stronge and Helm, Teacher Development by Iwanicki, and the Duties-Based by Scriven) deal with common elements such as the centrality of communication, need for specifications of expectations between evaluator and evaluatee, the need for establishing standards of performance, and provision for developing a professional development plan. Clearly, the paper presentations and flexibility in guiding discussions exercised by the seminar chair and CREATE's director made it possible to spend time discussing the four "promising" models and the potential for developing a common personnel evaluation model. It is recommended that active CREATE members and others interested in model building should devote some research effort to developing a common personnel evaluation model and that any future cross-cutting theory seminars should include a position paper dealing with key issues such as the development of a generalizable evaluation model.

It is difficult to convene a group to generate ideas and at the same time expect it to develop a concrete work plan for implementing those ideas. Another smaller group of those actively and intimately involved in CREATE's work would be in a better position to use the ideas generated at the conference to develop a plan. It is therefore recommended that future seminars should focus on generation of ideas. Position papers on the development of common evaluation models and/or concept papers should guide these seminars that will help tie together CREATE's work to maximize its impact on school personnel and school evaluations.

The seminar was less successful in developing a concrete plan for the cross-cutting theory. This view is confirmed by the proportion of participants who were somewhat uncertain about the success of the seminar in identifying and delineating issues and developing a work plan (38 percent, see Table 4). In large part this outcome is understandable because a group that is set up to generate ideas may not be an appropriate group to also undertake concrete planning of what should be done with the ideas. The seminar was a success in providing ideas that the Center can use in planning future activities. It is noted that although much emphasis in group discussions was on issues other than those specified in the seminar agenda, groups reviewed those issues and made some helpful suggestions. For example, groups agreed that CREATE needs to reach three distinct audiences: practitioners, evaluation research community, and policymakers at all levels, since these have the wherewithal to implement effective models and use the training kit and other products of CREATE. It was further suggested that CREATE should consider keeping a "warm body" in Washington, DC. Groups suggested that kits to be produced should help practitioners and school systems to conduct more feasible, useful, ethical, valid, and reliable evaluations. There were doubts about the utility of designing evaluation course materials for undergraduate teacher training programs, because it was argued that these programs already have too many courses. Essentially, the groups did not rule out any of the activities suggested in the agenda for the cross-cutting project and left it to CREATE to determine the priorities.

Participants were happy with the logistical support provided. They were pleased with their accommodations and meals at the Radisson Hotel. Thus, venues for future cross-cutting activities should provide similar support and facilities.

APPENDIX A

DISCUSSION DRAFT OF A TENTATIVE WORK PLAN FOR PROJECT 4.1 - CREATE STUDY COMMITTEE ON THEORY AND PRACTICE IN EDUCATIONAL EVALUATION: CROSS-CUTTING THEORY AND PRACTICE

Daniel L. Stufflebeam
Western Michigan University

DISCUSSION DRAFT OF A TENTATIVE WORK PLAN FOR PROJECT 4.1 - CREATE STUDY COMMITTEE ON THEORY AND PRACTICE IN EDUCATIONAL EVALUATION: CROSS-CUTTING THEORY AND PRACTICE

Daniel L. Stufflebeam[1]
Western Michigan University

BACKGROUND

In negotiating the 1992-93 continuation grant for CREATE, OERI required that CREATE complete a plan by the end of October 1993 for coordinating and integrating findings and products from CREATE's various projects. The purpose of the mandated cross-project work is to identify and address pervasive research and development issues and help educators to make maximal use of CREATE contributions for improving school evaluation practices. Thus, CREATE is currently planning a pertinent cross-cutting project for implementation during 1993-94 and 1994-95.

CREATE and OERI agreed that the main planning mechanism would be a working conference, i.e., the June 2-3, 1993, meeting in Kalamazoo. It will be attended by CREATE's project directors and National Advisory Panel and a group of distinguished practitioners and scholars. They will discuss the need and opportunities for addressing cross-cutting issues and integrating CREATE's research and development findings and products. They will also help plan the cross-cutting theory and practice project.

CREATE has reserved $111,723 in its proposed 1993-94 budget for implementing the proposed project's 1993-94 work and tentatively plans to budget a similar amount for the 1994-95 activities. CREATE has also projected that its National Advisory Panel will concentrate its fourth and fifth funding year efforts on helping CREATE staff to identify and address cross-cutting issues and integrate project findings and products.

OVERVIEW OF THIS PAPER

The purpose of this paper is to provide one input for the planning process that will occur at the June 3 and 4 cross-cutting theory and practice conference. In effect, this paper is a "placeholder plan" intended to help focus and guide discussion. By laying out a tentative plan for CREATE's cross-cutting theory development and product development work during the final 2 years of funding, I have sought to identify potential beneficiaries of CREATE's cross-cutting efforts, to hypothesize about what these stakeholders most need and would find useful from CREATE's efforts to integrate its work and package its contributions, to identify a limited range of integrative efforts that CREATE could pursue, to show alternative possibilities for CREATE's use of its available funds for this project, and to consider whether the cross-cutting project might be but the beginning of a longer-term professional exchange about how to improve teacher evaluation and other forms of educational evaluation.

---

[1] The example work plan for developing a sourcebook for educating teachers about educator evaluation, which appears as Appendix A of this paper, was developed based on a plan previously written by William Wiersma, James Sanders, and others.

While I have projected three possible lines of development to follow, I have not given the answers to the questions that must be addressed in developing our plan. Indeed, I am certain that I have not even identified all the right questions. But, I hope this paper will move us fairly directly to the issues we have to address and decide.

The remainder of this paper is a tentative plan for submission to OERI next October. It is organized to address the following questions:

1. What is the appropriate purpose of this project?
2. What is the project's relevance to CREATE's mission and other projects?
3. What projects should provide the primary source of information and products for this cross-cutting issues project?
4. What audiences can most benefit from this project and what do they need that CREATE might be able to provide?
5. What strategy should guide the work of this project during its remaining 2 1/3 years and what longer-range program should be considered?
6. What deliverables can reasonably be produced through implementation of the projected research and development strategy?
7. What set of participants should be called on to perform the work of this project?
8. What main tasks or subprojects should be performed by this project?
9. What schedule of work is required to fulfill the objectives of this project?
10. How should the available money and other resources be allocated?

1. What is the appropriate purpose of this project?

In view of the OERI mandate, the goal proposed for this project is to coordinate and integrate findings from CREATE and other Evaluation Center projects in order to improve theory and practice of evaluation as applied to U.S. schools. Explicitly included in this charge are (a) improving evaluations of teachers, administrators, support personnel, programs, and schools; (b) improving the instruction of educators for conducting personnel and school evaluation; and (c) focusing and motivating needed research and development in the area of educational evaluation. This project should help keep CREATE's researchers and advisors talking and thinking together about the relevant issues. It should be designed to uncover and explore ideas that emerge from individual projects, identify consistencies and commonalities across projects, resolve cross-project discrepancies, address development issues that are pervasive in CREATE projects, and purposefully integrate findings and products. This process will provide an "interaction effect," ensuring that the whole of CREATE is greater than the sum of its parts and reinforcing its emphasis on producing evaluation information and products that schools can incorporate into their policies and operations.

Beyond the immediate need to make the most of CREATE's other projects during CREATE's remaining 2 1/3 years, this cross-cutting project might appropriately address a longer-range need to foster dialectical exchange, on some regular and organized basis, among practitioners and scholars regarding the fundamental and intertwined issues in the evaluation of people, programs, and institutions.

There is certainly a need for such exchange, as illustrated by the provocative paper on credibility and politics of teacher evaluation prepared for this conference by Gene Glass and Barbara Martinez (1993). They raised the critical question of whether any steps to improve teacher evaluation can be effective in the climate of command that currently dominates teacher evaluation practices. Their argument to the contrary raises the further question of what, if anything, should and can be done to change the policies and political structures that govern teacher evaluation. In a similar vein, Jason Millman and Gary Sykes (1992), William Sanders and Sandra Horn (1993), William Webster, Robert Mendro, and Ted Almaguer (1993), William Webster and Marvin Edwards (1993) and others are pursuing serious research and development on the familiar question of whether school districts, schools, and teachers can be fairly evaluated based on the test scores of their students. This work has definite policy implications as is seen in the statewide tests of the idea to be conducted in Tennessee over the next three years. These examples of current inquiry into basic questions about educational evaluation reinforce the idea that school policy development efforts could benefit by ongoing disciplined discussion of the perplexing and fundamental issues that educators must thoughtfully and productively address over an extended period if schools are ever to change and improve their evaluation practices.

If we desire it and if we pursue it, this cross-cutting theory project could be used to spawn ongoing periodic exchanges among scholars and practitioners about the basic issues in educational evaluation, particularly teacher evaluation. One example of how such an ongoing exchange might develop was seen in the "May 12 Group" (so called because it met for the first time on May 12, 1969). For about 25 years this relatively small group, with an evolving membership, met at various places to debate issues and outline ideas for improving the theory and practice of program evaluation. This group had important exchanges about and did influential work on such notions as the definition of evaluation, the limitations of experiments for evaluating programs, the nature and role of professional standards for evaluation, and goal-free evaluation. So far as I know, there was never any budget to support this ongoing exchange, only the personal or institutional funds that each participant applied to their participation. If it seems important to consider whether a "May 12-type group" should be developed, for example, for the area of teacher evaluation, happily, our conference could consult with several of our cross-cutting project participants who had extensive experience with the May 12th Group. They include David Berliner, Gene Glass, Richard Jaeger, and Michael Scriven.

In any case, I am mindful that our first purpose in designing the cross-cutting project should be to fulfill our contractual obligation to OERI to produce useful integrations of CREATE findings and products. In addition, this project may give us a valuable opportunity to start up a sustained, disciplined exchange about pervasive issues in evaluating teachers and other aspects of education, extending beyond the funding period for CREATE.

2. What is the project's relevance to CREATE's mission and other projects?

CREATE's mission is to conduct research, development, and dissemination to assist U.S. school districts to (1) validly and reliably evaluate professional educators and institutional performance, (2) effectively use the evaluation results to improve educational services to students and communities, and (3) credibly and systematically assure accountability to constituents and sponsors.

Project 4.1 is intended to address all three of these objectives. It is directed at compiling CREATE findings and products into theoretical contributions that can help focus research efforts and provide guidance for improving evaluation practice; sourcebooks that can be easily accessed and used by those involved in educating teachers and other educators; and kits that school personnel can use to assess and improve teacher evaluation practices. Through focused research and development and effective use of the sourcebooks and kits, improved educational service to students and communities and improved accountability are intended.

The resources for this project are obviously limited, and it is probably inadvisable to concentrate the cross-cutting effort on all aspects of CREATE's work. Considering that about 60 percent of CREATE's budget is devoted to teacher evaluation, I have directed the tentative plan being presented in this paper to the topic of teacher evaluation. At the same time, I recognize that teacher evaluation around the country has low credibility, partly because school administrators typically are not evaluated in any depth. This dilemma deserves consideration as we focus the cross-cutting project.

3. What projects should provide the primary source of information and products for this cross-cutting issues project?

The CREATE programs and associated CREATE and Evaluation Center projects (past and present) on which this integration effort can usefully draw are as follows:

Program 1.0: Improvement of Teacher Performance Evaluation

- Teacher Evaluation Models Project (Scriven: 11/90 - 10/93)
- Improved Teacher Evaluation Models Development (Scriven: 11/93 - 10/95)

- Grounded Theory of Teacher Performance Evaluation (Stufflebeam: 1/92 - 10/93)

- Development of Classroom Assessment Techniques for Teacher Self-Evaluation (Gullickson/Airasian: 5/92 - 10/94)

- Expert Science Teacher Project (Burry: 11/90 - 10/93)

- National Fast Response Survey of Teachers (Stufflebeam: 12/91 - 12/93; funded by NCES, not directly a CREATE project)

- National Survey of Schools (Barley: 11/90 - 10/91)

Program 2.0: Improvement of Administrator and Support Personnel Evaluation

- Models for Administrator Evaluation (Stufflebeam/Bridges/Candoli/Cullen - 1/93 - 10/95)

- Model for Evaluation of Professional Support Personnel (Stronge/Helm: 11/92 - 10/95)

Program 3.0 Improvement of School Evaluation

- Improved and Tested School Evaluation Models (Sanders: 11/93 - 10/95)

- Development of a Research-Based School Report Card (Nyirenda/Jaeger: 05/92 - 10/93)

- Models for School Evaluation (Gallegos/Benjamin/Candoli/Wegenke: 11/90 - 10/92)

Program 4.0 Theory Development and Special Projects

- CREATE Study Committee on Theory and Practice in Educational Evaluation: Cross-cutting Issues and Practice (Stufflebeam/Ervin/Gullickson: 11/92 - 10/95)

- Adapting W. Edwards Deming's Statistical Process Control Model (Bayless/Massaro: 1/91 - 10/92)

- Evaluating the Ability to Work With At-Risk Students and Their Families (Lavely/Blackman: 11/90 - 10/92)

Program 5.0 National Evaluation Resource Service

- Dissemination of Research and Products (Gullickson/Ervin: 11/90 -10/95)

- National Evaluation Workshops (Sandberg/Keller: 11/91 - 10/95)
- CREATE Newsletter (Ervin: 11/90 - 10/95)
- Evaluations of Nationally Significant Evaluation Programs, particularly the NAGB Levels Project (conducted by Jaeger, Scriven, and Stufflebeam, 1990 and 91)

One of the inputs being developed for this conference is a paper by Gullickson outlining the issues that cut across CREATE's projects and examining the perspectives, approaches, and findings that relate to the common issues. The Gullickson paper and project summaries in Part A of the continuation proposal provide a relevant data base for planning the cross-cutting project.

4. Which audiences can most benefit from this project and what do they need?

This project is tentatively designed to serve four principal audiences. First, the project would serve CREATE's project directors and staff in clarifying CREATE's approach to cross-cutting issues, e.g., guidelines for developing evaluation manuals that lead to change in practice, a common approach to developing user-friendly consumer reports, a concerted approach to studying and addressing political issues in evaluations, and a shared view of the requirements of sound theory. Second, we intend to develop and package materials in kits that are directly aimed to assist practitioners in schools (teachers, administrators, and school board members) to apply CREATE's findings and products to assess and improve their evaluation practices. Third, we would develop sourcebooks for use by teacher educators to provide both preservice and inservice education in evaluation to teachers. Fourth, we plan to package and disseminate CREATE findings and products for use by those outsiders who work in behalf of schools, e.g., researchers, state and federal policy officers, evaluators, philanthropists, and professional association officers.

5. What strategy should guide the work of this project during its remaining 2 1/3 years?

The primary activity of this project for 1992-93 is to develop a plan for the remaining two years of the project. The project is being launched by means of a seminar. We think it will be useful to continue to use the seminar format for periodic review of progress, presentation and discussion of issue papers, and planning of future project activities. Between seminars, staff from this project would use CREATE and other Evaluation Center project reports, instruments, and related materials to compile findings and synthesize information into working documents (monographs, guidelines, kits, sourcebooks, etc.). These would then be presented, discussed, evaluated, and modified at one working seminar each year. Participants would include members of the National Advisory Panel, CREATE staff, and others as needed. As presently envisioned, the seminar will occur annually as a preface to the annual National Advisory Panel meeting. This format would enable Panel members and project directors to have substantive input and would both foster development of integrative approaches to evaluation and provide direct feedback to individual project directors and staffs.

6. What deliverables and other outcomes can reasonably be produced through implementation of the projected research and development strategy?

We expect two kinds of outcomes from this project, one formative in nature, the other summative. The first is feedback--ideas and information to help CREATE directors and staff focus and conduct their individual projects. The second is published products designed to guide evaluation practice and evaluation training and to focus future inquiry on pervasive and critically important issues.

Through this integrative-approach, the project will help CREATE to develop

a. issue papers and monographs

b. sound ways of addressing issues that cut across CREATE's research and development plus tangible illustrations of how CREATE addressed the issues

c. a cohesive set of supporting materials to guide educators in their conduct of teacher evaluation practices, e.g., guidelines, example instruments, lists of criteria, sample board policies

d. sourcebooks for use in preparing teachers to be proficient in conducting and using evaluations

e. a comprehensive approach to developing theory for use in understanding and guiding evaluation practice

7. What set of participants should be called on to perform the work of this project?

This project will directly involve virtually all CREATE project directors, core staff members, and Advisory Panel members, as well as a group of distinguished scholars and practitioners.

8. What main tasks or subprojects should be performed by this project?

Pursuant to the purposes addressed above, and as funds permit, it would be appropriate for this project to carry out at least the following tasks:

a. Development and dissemination of a series of papers and monographs that address pervasive issues in evaluation, e.g., the papers that were prepared for the project's planning seminar by Glass, on the politics of evaluation, and by Scriven, on the nature of sound and useful theory contributions. Other papers might focus on the requirements of sound and useful manuals for implementing evaluation models, sound and useful consumer evaluation reports, and analyses of the efforts to use student performance measures to evaluate teacher performance.

b. Development and dissemination of sourcebooks for use in preparing teachers to plan, conduct, assess, and use evaluations. A two-year plan has been developed for this subproject, as presented in CREATE's Year 4 continuation proposal and as summarized later in this paper.

c. Development of a teacher evaluation improvement kit that schools could use to assess and strengthen or replace their present teacher evaluation systems. A prospectus for this proposed kit may be found in Attachment 3. Such a kit could include the Joint Committee *Personnel Evaluation Standards*; a diagnostic instrument, from CREATE's Evaluation Theory Development Project, for schools to use in analyzing and assessing their present evaluation system; sample evaluation policies that adhere to the Joint Committee Standards; a manual for organizing an evaluation system improvement project; Scriven's "Duties of the Teacher" and consumer report on alternative evaluation models; and brief manuals for addressing particular tasks in evaluation work (e.g., writing and keeping position descriptions up-to-date, taking student achievement into account in evaluating teacher performance, guidelines and pitfalls in merit pay evaluation, remediating and terminating ineffective teachers).

9. What schedule of work is required to fulfill the objectives of this project?

The actual work plan and the schedules for this project during Years 4 and 5 will depend upon the results of the planning seminar and follow-up project work during Year 3 of CREATE. Thus, we cannot yet provide a finalized work plan nor the schedule for Years 4 and 5. During September Drs. Stufflebeam and Gullickson will provide, as this project's 1992-93 deliverable, the substantive planning document for Project 4.1's work during the 1993-94 and 1994-95 project years.

As an illustration of the detail required in our work plan for this project, one possible subproject for this project has been projected by Wiersma, Sanders, and others: the development and field testing of curriculum components for training teacher education students in educator evaluation. This possible subproject is described in Attachment 1 to illustrate the kind of work to be done through Project 4.1. Attachment 3 contains a working outline for delineating Project 4.1 and subprojects.

10. How should the available money and other resources be allocated?

Approximately $110,000 is projected to be available for the Cross-cutting Issues Project during each of the remaining CREATE grant years (1993-94 and 1994-95). As a general guideline, each year $38,000 could be allocated to seminars and preparation of issue papers and monographs, $44,000 to the development of two sourcebooks for training teachers in educator evaluation, and $38,000 to the development of a teacher evaluation improvement kit. The above only indicates the general ballpark for the effort. A more definitive allocation must await adjudication and delineation of the work plans. Available options include but are not limited to assigning all of the budget to two subprojects, or only one.

REFERENCES

Glass, G. V., & Martinez, B. A. (1993). *Politics of teacher evaluation.* Unpublished manuscript.

Millman, J., & Sykes, G. (1992). *The assessment of teaching based on evidence of student learning: An analysis.* (Research Monograph No. 2). National Board for Professional Teaching Standards.

Sanders, W. L., & Horn, S. P. (1993). *The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment.* Manuscript submitted for publication.

Webster, W. J., & Edwards, M. E. (1993, April). *An accountability system for school improvement.* Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

Webster, W. J., Mendro, R. L., & Almaguer, T. O. (1993, April). *Effectiveness indices: The major component of an equitable accountability system.* Paper presented at the meeting of the American Educational Research Association, Atlanta, GA.

ATTACHMENT 1

ILLUSTRATIVE PROJECT 4.1 SUBPROJECT: DEVELOPMENT OF SOURCEBOOKS FOR TRAINING TEACHERS IN EDUCATOR EVALUATION[2]

[Contingent on the outcomes of CREATE's June 1993 Cross-cutting Theory Seminar, the sourcebook development subproject is planned for November 1993 - October 1995.]

ABSTRACT

This subproject of the cross-cutting project is designed to develop the curriculum component for teaching about educator evaluation in teacher preparation programs or inservice programs.

The products are projected to be two field-tested sourcebooks containing personnel evaluation materials to be used in training programs for educators: one for preservice use and one for inservice use. The intended outcomes of this subproject of Project 4.1 are to prepare instructional materials on the concepts and procedures for evaluating school personnel. The sourcebooks would consolidate field tested essential information on which to base evaluations of school teachers, principals, superintendents, and support personnel.

INTENDED BENEFICIARIES AND ASSESSMENT OF THEIR NEEDS RELATED TO THIS PROJECT

This effort would collaborate with and provide assistance to

- college and university instructors in teacher preparation programs
- inservice and staff development directors and instructors

The audiences impacted are teachers, both prospective and inservice, teacher educators, and evaluators of teachers, all of whom would benefit from an evaluation component in the curriculum and inservice training programs focusing on evaluation.

STATEMENT OF THE PROBLEM TO BE ADDRESSED

According to the findings of the Teacher Evaluation Models Project (TEMP), schools of education do not traditionally teach prospective educators about teacher evaluation [gaps in the information will be addressed in the supplementary literature review for this subproject]. The quality of educational evaluation could be substantially improved if preservice teachers received a thorough introduction to evaluation as a component of their curriculum, and if inservice teachers received this training during their tenure.

---

[2] Based on a plan by William Wiersma, James Sanders, and others.

GOALS AND OBJECTIVES

- to develop the curriculum component for teaching about educational evaluation in educator preparation programs or inservice programs
- to compile two sourcebooks for preservice and inservice programs
- to field test the sourcebooks in undergraduate and inservice educational settings
- to disseminate the field-tested sourcebooks to higher education faculty and school personnel administrators through publications and conference presentations

RELATIONSHIP TO CREATE MISSION AND CURRENT PROJECTS

Sourcebook development activities follow directly from the purposes and intended outcomes of Project 4.1 and build directly on the work of CREATE projects. This subproject would begin in Year 4, at which time findings from CREATE's projects on teacher evaluation and administrator and support personnel evaluation will be available as resources for this project.

PRIMARY SOURCES OF INFORMATION AND PRODUCTS FOR THIS CROSS-CUTTING ISSUES PROJECT

Primary sources of information for the subproject would be relevant materials from CREATE projects and a supplementary literature search, plus feedback from field tests and reviews of the sourcebooks.

METHODOLOGY

A primary task of this sourcebook development effort is to synthesize and organize other projects' findings into readily usable instructional materials that can be applied as instructional modules in undergraduate and graduate professional education programs and as stand-alone support materials for use in inservice education. A key concern is the preparation of materials that effectively deal with the issues but that are sufficiently brief and self-contained that they can be easily incorporated into the above noted instructional situations.

Project 4.1 staff would collect materials for teacher training in personnel evaluation from other CREATE projects and compile them into the sourcebooks. These sourcebooks would be reviewed by CREATE's National Advisory Panel and subsequently discussed in a meeting of the Panel. Staff would make indicated changes, then field test the sourcebooks in undergraduate, graduate, and inservice education settings. Findings would be discussed with CREATE's NAP and also communicated to higher education faculty and school personnel administrators through publications and conference presentations.

This would be a two-year effort beginning with CREATE's Year 4. The first year's work would concentrate on developing the sourcebooks and planning the field tests. Because the CREATE project year does not coincide with the university academic year, part of the field test might begin during the final two months of the first project year. Initial work on the sourcebooks would also be done in the first year.

In Year 4 there would be an initial review of relevant materials from related CREATE projects and a supplementary literature search to provide content and structure for the sourcebooks. The available information would be organized and synthesized into an evaluation component for each training curriculum. The evaluation component would be in a form ready for field tests in training programs and inservice workshops. Developing the sourcebooks would involve sorting and synthesizing the results from related CREATE projects (as identified above). The sourcebook's contents would be selected according to the needs of both inservice and preservice teachers. Systems for categorizing materials would be developed. Syllabi would be requested as available, and a common syllabi format would be adopted. The syllabi would be prepared in draft form and circulated for review by CREATE's NAP and prospective sourcebook users.

The second year would be devoted to completing the field tests and revising and finalizing the sourcebooks. It should be noted that the field tests would be conducted in undergraduate programs as well as inservice workshops. As time allows, project dissemination would be started in the second project year and would become part of CREATE's national dissemination effort. Part of that effort would be to publish the curriculum components from each sourcebook in outlets regularly read by teachers, principals, superintendents, support personnel, and relevant higher education faculty.

The subproject would consider options for incorporating the evaluation component into existing training programs. The evaluation component would probably be included in the "educational foundations" portion of the undergraduate program. Colleges of education would have options in adapting the component for their own programs. Most institutions would probably take the option of including the component in an existing course or part of a program. However, a separate course would also be an option, and it might also be desirable to incorporate at least part of the instruction in a professional field experience.

The preservice field tests of the sourcebooks would be conducted in about five settings for each sourcebook aimed at getting a diversity of training contexts. Selected characteristics of desired settings include (1) a state-supported university with large educator programs and a semester system, (2) a state-supported university with large education programs and a quarter system, (3) a large private university with large educator programs and a semester system, (4) a midsized liberal arts college with substantial educator programs and a semester system, and (5) a small liberal arts college with educator programs on the semester system. Inservice workshops would be planned for large, middle-sized, and small school districts in at least three states. Programs in these various institutions should provide for good contrast for the field test and permit credible research results. The draft of the component would be circulated to all field-test sites and to CREATE's NAP, followed by a meeting of the project team. At this meeting the field-test draft would be finalized along with the field test plans.

DELIVERABLES AND DUE DATES

The deliverables would include (1) a draft of the evaluation components in preparation for the field tests, (2) the evaluation curriculum components and accompanying materials appropriate for both preservice and inservice/staff development programs, and (3) the two sourcebooks.

DISSEMINATION PLAN

The subproject's concluding activity would be to begin dissemination. CREATE's national mission and dissemination capabilities gives it a unique opportunity to provide extensive dissemination to the profession, not only to teacher education institutions but also to elementary and secondary schools. Dissemination to such schools would be for informational purposes, especially concerning inservice offerings.

The American Association of Colleges for Teacher Education (AACTE) could be especially helpful in disseminating to teacher education programs because of its extensive membership and high status in the profession. AACTE's involvement as a cosponsor or strong supporter of the sourcebook development effort would enhance the perceived relevance and acceptance of CREATE materials, especially the evaluation component's materials.

Publications in outlets read by teachers, principals, superintendents, or support personnel would be prepared during Year 5.

EVALUATION PLAN

The field tests, as described in the methodology section, would be conducted in accordance with the guidelines of the *Handbook for Review and Validation of CREATE Projects*. The process would include evaluation of the field tests and project team meetings for revision, exchanges/meetings with the CREATE NAP and finalization of the evaluation component.

YEAR 1993-94 AND YEAR 1994-95 TIMELINE

Development of the sourcebooks is projected to require 24 months commencing on November 1, 1993, and concluding on October 31, 1995. The time line for the tasks is as follows: (1) reviewing the plan for this subproject with CREATE's NAP and selected advisors (6/93), (2) conducting a comprehensive literature search and compiling CREATE findings related to each component (11/1/93-2/28/94), (3) developing the sourcebook for each curriculum component (3/1-7/30/94), (4) reviewing progress with CREATE's NAP (7/94), (5) planning/organizing the field tests (5/1-8/31/94), (6) conducting the field tests (9/1/94-2/28/95), (7) analyzing field tests results

(3/1-5/31/95), (8) reviewing field test results with CREATE's NAP (6/95), (9) revising and finalizing each sourcebook (6/1-7/31/95), and (10) disseminating products (8/1-10/31/95).

KEY STAFF

The sourcebooks would be developed by a task group to include Ms. Rebecca Thomas (or another Evaluation Center staff member) and Dr. William Wiersma, University of Toledo. They would be assisted by the National Advisory Panel and by consultants, e.g., Dr. Robert Gafka, Defiance College, and Dr. Stephen Jurs, University of Toledo.

BUDGET

[To be determined.]

ATTACHMENT 2

DRAFT OUTLINE FOR THE PROPOSED CROSS-CUTTING THEORY AND PRACTICE PROJECT

The following are tentative outlines for an overview of the cross-cutting theory and practice project and for each subproject that might make up the overall project. Obviously, the plan for each subproject must be completed before the overview can be finalized. In general, these outlines include the items that OERI has asked us to address.

At this point, I am thinking that CREATE might pursue up to 3 subprojects: a theory seminar focused on cross-cutting evaluation issues, a subproject to develop sourcebooks for use in educating teachers in the design and use of evaluation, and another subproject for schools to use in analyzing and strengthening their teacher evaluation systems. These are ideas to get us started. In view of the budget limitation, it may be unrealistic to undertake this many subprojects, and you may think that some other topics are more important. A key task for each working group at the June 2 and 3 Cross-cutting Theory Project meeting is to advise whether there should be one cross-cutting project or several subprojects and what topic(s) would be most appropriate to address.

For each item in the following outlines I have included notations giving my ideas about what is most important for you to consider and what leads, ideas, and CREATE information you might draw upon. CREATE will have to fill in the complete details before it submits the planning document to OERI next October. I understand that in a 2-day conference, we can only do a part of what needs to be done. Below, I have tried to direct your attention to the items where we most need your advice.

PROJECT OVERVIEW

Purpose of the Cross-cutting Theory and Practice Project (Please review and improve as you see fit the purpose statement as presented at the outset of this paper; also incorporate the Scriven paper/presentation on the nature, requirements, and uses of sound theory.)

Project's relevance to CREATE's mission and other projects (Please review and improve the statement of relevance to CREATE's mission and what you see as desired interconnections between this project and CREATE's other projects.)

Potential beneficiaries and analysis of their needs that CREATE might address through this project (I have had in mind evaluation researchers and their needs for analysis of assumptions, issues, and new findings underlying the practice of teacher evaluation; teacher educators and their need for teacher training materials focused on the design and use of evaluation; and school district and building personnel and their need for clear, practical direction for diagnosing and strengthening teacher evaluation policies and practices. Should this view be narrowed, broadened, made more specific, or what?)

Analysis of relevant cross-cutting issues seen across projects (This is a priority item for your attention. I suggest that you look at Part A of the 1993-94 CREATE continuation proposal, Gene Glass's paper on the politics of evaluation, and Arlen Gullickson's paper on themes and findings seen in CREATE's projects. Two issues here in Michigan that might be candidates for attention include achieving sound teacher evaluation in the context of collective bargaining agreements and the joint and individual roles of school districts and the state in using teacher evaluation to protect the interests of students.)

Projects and other sources that provide the primary information and products for use by this cross-cutting theory and practice project (Two key sources for you to review here are Part A of CREATE's 1993-94 continuation proposal and the Gullickson paper for this conference. Also, CREATE's Edith Ervin and Stanley Nyirenda are key resource persons concerning respectively past project reports and documents and evaluation of CREATE's projects and products.)

Overall project strategy for 1993-94 and 1994-95 (Generally, is it a good idea to use the seminar format to plan and guide the cross-cutting effort? Should we also support cross-cutting issues subprojects, such as a teacher education evaluation sourcebook subproject and a teacher evaluation improvement kit subproject?)

Main tasks or subprojects (This is the fundamental question to be answered. Should there be one overall project or several subprojects? What main outcomes should be pursued?)

Main deliverables and due dates (What main items do you think this project should produce at the end of each project year?)

General schedule of work (If there is to be a NAP seminar each year, when should it occur, what type of staff work should be done before the seminar, and what should be done to follow up the seminar, e.g., a proceedings report? What main milestones should NAP expect from the other subprojects, if there are such? This part of the overview must await the completion of plans for the individual subprojects and thus is not a priority item for the 2-day cross-cutting issues meeting.)

Dissemination plan (General ideas would be welcomed, but this is not a priority item for the 2-day meeting.)

Evaluation plan (Dr. Nyirenda is in charge of preparing this part of the plan. I am sure he would welcome your suggestions. But this is not a priority item for the 2-day cross-cutting issues meeting.)

Staff and other participants (What we need here, especially, are specific suggestions where individual NAP members and others desire to be involved in the cross-cutting issues work, e.g., writing issue papers, developing products, helping with review and field testing, assisting dissemination.)

Guidelines for allocating the projected $220,000 for this project, assuming a 20 percent indirect cost factor (Basically, on what topics/subprojects do you think our available funds can best be spent and what pattern of allocation would be appropriate?)

Consideration of a longer-range program (Some limited discussion should be devoted to the question of long-range continuation of the work started in this project, but this is not a priority item for the 2-day meeting.)

OUTLINE FOR EACH SUBPROJECT

Abstract (As one of the last tasks, complete a one-paragraph abstract of the subproject--not a high priority item, but CREATE staff would welcome your submission.)

Intended beneficiaries and assessment of their needs related to this subproject (Basically, what group of professionals would most benefit from this subproject, and which of their needs do you think the subproject should target? This is a high priority item for your attention.)

Statement of the problem to be addressed (Please address this high priority item by defining as precisely as you can the problem that reasonably should be targeted during the 2-year subproject.)

Goals and objectives (Give high priority to listing the subproject objectives.)

Relationship to CREATE mission and current projects (Give a succinct justification for this subproject in the context of OERI's request for cross-cutting work and of CREATE's mission and other projects.)

Primary sources of information and products for this cross-cutting issues subproject (Not a high priority item for the 2-day meeting. Nevertheless, your suggestions about what CREATE should do to support this subproject through its other projects and activities would be welcome.)

Methodology (This high priority item requires that you lay down the basic method to be followed in the subproject. Please feel free to give specific requirements that CREATE should meet in carrying out the subproject.)

Deliverables and due dates (This is a high priority item. Please give specifics.)

Dissemination plan (This is a moderate priority item. Your general suggestions will be welcome.)

Evaluation plan (Please list the main evaluation questions that should be addressed in evaluating the success of this subproject. Further suggestions about evaluation methods are welcome but not at this time a high priority, since Dr. Nyirenda will be working out the details.)

Year 1993-94 time line (A general list of tasks, assignments, and time spans would be helpful, but mainly to help assure yourselves and CREATE staff that your proposal is feasible and to help us consider its feasibility when grouped with other possible subprojects.)

Year 1994-95 time line (A general list of tasks, assignments, and time spans would be helpful, but mainly to help assure yourselves and CREATE staff that your proposal is feasible.)

Key staff (General suggestions are welcome, but this is not a high priority item for the two-day meeting.)

Budget (Please do enough so that you and we know about how much this subproject would cost.)

ATTACHMENT 3

TEACHER PERFORMANCE EVALUATION SYSTEM IMPROVEMENT KIT
A Prospectus

The *Teacher Performance Evaluation Improvement Kit* is projected to be a practical, step-by-step guide to

- assessing and reporting on the strengths and weaknesses of an existing teacher performance evaluation system
- designing a new or improved teacher performance evaluation system
- formulating appropriate policy to authorize and guide the use of the teacher performance evaluation system
- training the evaluators
- assessing and reporting on teacher performance
- Using teacher performance evaluation results both to help improve teaching and to help assure teacher accountability

The projected KIT is intended for use by school boards, administrators, teachers, teacher and administrator educators, and evaluation consultants. It would be written by researchers and practitioners with substantial experience in the schools. It would be designed to serve as a useful guide for generalists, as well as a practical and comprehensive set of references for experienced evaluation administrators, trainers, and consultants.

The KIT would be grounded in the Joint Committee *Personnel Evaluation Standards*, which is intended to be the major authoritative statement by professional and lay groups, concerned with education, regarding what constitutes sound and useful personnel evaluation in U.S. schools. The KIT also would incorporate the research findings and products produced by the U.S. Department of Education-funded national Center for Research on Educational Accountability and Teacher Evaluation (CREATE).

The KIT would lead the user step-by-step through the entire teacher performance evaluation improvement process, from assessment of the existing evaluation system through design, policy-writing, training, implementation, and use of findings. The user-friendly format would include brief manuals designed collectively to guide the users through the complete teacher evaluation improvement process, or to serve as focused references for use in addressing particular evaluation tasks. The manuals would contain essential definitions and rationales; step-by-step processes; checklists of helpful hints and pitfalls to avoid; example job descriptions, data collection forms, diagnostic charts, and report formats; plus many other procedural items. Users could purchase any or all of the manuals in the kit, which would total about 350 pages.

Tentatively, the KIT would contain the following modules:

1. What Professional Standards Should Undergird Teacher Evaluations? (an approximately 20-page manual defining the Joint Committee Personnel Evaluation Standards in relationship to assessing and designing teacher evaluation systems)

2. How Can a District Effectively Use Professional Standards to Evaluate a Teacher Performance Evaluation System? (a step-by-step illustrated manual, approximately 20 pages, providing a process for examining a teacher performance evaluation system with special attention to political issues. It would draw on the CREATE Grounded Theory of Teacher Performance Evaluation Project; the Reineke, et al. article on the experience of the Lincoln, Nebraska, Public Schools in using the Joint Committee Standards to assess and revise the district's teacher evaluation system; and Part II of The Personnel Evaluation Standards on how to apply the Standards).

    2.1 What's the step-by-step process for assessing and improving or replacing a teacher performance evaluation system?

    2.2 How can a district develop and maintain credibility for its teacher performance evaluation system, including, importantly, the selection and training of personnel? (Especially, who should be involved, what decision rules are appropriate for improving the system, how can stakeholders make suggestions or file grievances, how will the evaluators be selected and trained, how can the evaluators be held accountable for maintaining utmost professionalism, how and with what frequency should the performance evaluation system be formally reviewed, what can be done to assure that the evaluations contribute to improvements in teaching?)

    2.3 What approaches are appropriate and effective for identifying and addressing political issues in teacher performance evaluation? (Identify and address the pertinent conflicts of interest and assure openness and equity in deliberations and decision making.)

3. What Should a District Include in a Sound Plan for Teacher Performance Evaluation? (an illustrated, approximately 40-page manual that outlines the elements of a sound teacher performance evaluation system; addresses specific substantive, legal, political, technical, and administrative issues to be resolved in formulating a district policy and plan; and provides sample policies for a teacher performance evaluation system)

    3.1 What points should be addressed in the district manual for teacher performance evaluation?

3.2 What are appropriate alternative/complementary purposes for teacher evaluation? (drawn from Stufflebeam's analysis of six alternative purposes for teacher evaluation)

3.3 What serious flaws are often seen in the prevalent practices of teacher evaluation? (drawn from Scriven's analysis of evaluation models)

3.4 What are the elements of a sound career ladder for teachers? (drawn from the career ladder experience in Tennessee)

3.5 How can a district assure that its teacher performance evaluation policy is consistent with pertinent state tenure laws and teacher evaluation policies? (review and discuss a representative set of state teacher evaluation policies)

3.6 How can teachers, board members, and administrators assure integrity of a teacher performance evaluation system that is generated through collective bargaining? (For example: don't include the criteria and instruments in the contract; do define teacher and evaluator roles; adopt the Personnel Evaluation Standards as the basic guiding policy.)

3.7 What is a defensible district process for remediating and/or discharging poorly performing teachers? (Draw from the works of Hans Andrews, Edwin Bridges, and Barry Groves)

3.8 What are the issues in formulating a defensible policy and procedure on merit pay? When is it appropriate to provide merit pay for groups of teachers rather than individuals and how can this best be done? (Draw from the work of William Webster)

3.9 What are the issues in formulating a defensible policy and procedure on incentive pay? (Draw especially from the work of Richard Benjamin)

3.10 What special provisiohs for teacher performance evaluations are required at the elementary, middle, and secondary school levels? (Draw on the works of Burry and Scriven)

4. What are Some Useful Resources for Clarifying Teaching Expectations? (an approximately 15-page booklet providing examples and practical advice)

4.1 Scriven's Duties of the Teacher

4.2 Advice on developing sound position descriptions and keeping them up-to-date (Draw from the relevant Joint Committee standard, A-2 -- Defined Role, and give examples.)

4.3 How to delineate the criteria and weights for evaluating teacher performance (Draw from experiences with the Hay System for classifying and evaluating the performance of personnel)

4.4 What are some example teacher job descriptions, cutting across school levels and content areas, that provide sound bases for performance evaluations?

5. How Can District Personnel Defensibly and Usefully Measure Teacher Performance? (set of focused manuals)

5.1 How to Help Teachers Assess Their Own Progress (about 20 pages; draw especially from the research of Airasian and Gullickson)

5.2 How to Obtain Student Input for Use in Evaluating Teacher Performance (approximately 10 pages; draw especially from Scriven's writings)

5.3 How to Plan and Conduct Peer Evaluations (approximately 20 pages; draw especially on the experiences of the school districts in Poway, California; Northfield, Minnesota; and Toledo, Ohio)

5.4 How to Develop and Use a Sound Teacher Evaluation Portfolio (about 15 pages; draw on the work of the National Board for Professional Teaching Standards)

5.5 How to Observe, Document, and Analyze Classroom Teaching (approximately 20 pages; draw on the works of Judy Burry, Madeline Hunter, and Richard Manatt)

5.6 How to Assess Teacher Performance Based on Student Outcomes, and What Pitfalls to Avoid (about 20 pages; draw on the works of Jason Millman, William Sanders, and William Webster)

5.7 How to Use Videotapes in the Process of Analyzing and Improving Teaching (about 20 pages; draw on the work of the National Board for Professional Teaching Standards)

5.8 How to Interview Teachers (about 15 pages; draw on the work of Eder and Ferris, also Martin Haberman)

6. How Can District Personnel Effectively Report and Apply Teacher Evaluation Results? (A set of focused manuals)

6.1 How to Conduct Feedback Sessions (about 20 pages; draw on the works of Madeline Hunter and Richard Manatt)

6.2 How to Develop Growth Plans (about 15 pages; draw on the works of Madeline Hunter and Richard Manatt)

6.3 How to Use Teacher Performance Evaluation to Focus and Guide Professional Development (approximately 10 pages; draw on the works of Madeline Hunter and Richard Manatt)

7. How Can a District Effectively Schedule and Manage Teacher Performance Evaluations? (approximately a 30-page management manual)

7.1 What is a reasonable annual calendar for teacher performance evaluation? (draw on the work of James Stronge and Virginia Helm)

7.2 Who should be in charge of making the evaluation system work?

7.3 How can the district efficiently and effectively train the evaluators?

7.4 What's a reasonable approach to budgeting for teacher evaluation?

7.5 What are some concrete suggestions for detecting and controlling bias in the teacher performance evaluation system?

7.6 What's a reasonable timetable and approach to reviewing and revising the teacher performance evaluation system?

8. How Can a District Apply the Teacher Performance Evaluation System to Particular Purposes? (A set of focused manuals)

8.1 How can a district effectively design and use evaluation to select competent teachers? (approximately a 15-page manual; draw on the work of Martin Haberman)

8.2 What is an effective approach to evaluating probationary teachers? (approximately a 15-page manual; draw on the work of the Toledo, Ohio, school district)

8.3 What is an effective approach to evaluating teachers for making tenure decisions? (approximately a 15-page manual)

8.4 What is an effective approach to evaluating and remediating poorly performing teachers? (approximately a 15-page manual)

8.5 What is a proven process for conducting and using evaluation to terminate persistently ineffective teachers? (approximately a 15-page manual to be drawn from the works of Hans Andrews, Edwin Bridges, and Barry Groves)

APPENDIX B

A CREATE OVERVIEW

Arlen Gullickson
Western Michigan University

A CREATE OVERVIEW

Arlen Gullickson
Western Michigan University

It is my task to provide you with an overview, a cross-cut if you will, of the CREATE R&D effort. I know that you already have in hand our proposal for Year 4 of CREATE. As you know, Section A of that proposal provides project-by-project summaries of CREATE efforts to date. These summaries were developed in response to OERI's requirement that we provide a description of project efforts and findings as the first major section of our proposal for continued funding. Additionally, Section B of the proposal contains our current plans for the future. Those two sections provide you with a fairly comprehensive perspective of what CREATE is currently doing and what it intends to do in the coming year.

This morning I will try to summarize the ideas, plans, and findings of the projects (those already completed, those currently underway, and those planned for the coming year). Hopefully, my presentation will help you better understand the larger intentions that we have for CREATE; that is, how the projects are woven together to provide fabric out of the single and multiple strands of the individual projects. In describing CREATE, I hope to refresh your memories about the projects themselves, the people engaged in the projects, what they are doing, and their intended products. I want to both synthesize information from the individual projects and to raise questions or issues that appear to be pertinent across the range of projects.

Much of what I have to say is already available to you. Some of you have read all of the documents that CREATE has produced and come to this meeting well grounded in CREATE. Every one of you has experiences well beyond the bounds of CREATE that provide knowledge, ideas, and perceptions that supplement the ones that I will present.

Hopefully, when I am done, you will have questions for project staff, issues to raise, and ideas both about the cross-cutting concerns and about the projects individually. We want you to know as much as possible about CREATE, because we think the more you know about us the better you will be able to help us as we tackle this question of what the cross-cutting project should be about in Year 4 of CREATE. This presentation will only be successful if it stimulates your ideas and provides a basis for thoughtful discussion and movement toward development of a strong plan of action.

With that preface, let me begin. First, let's quickly review CREATE's stated mission. In the original proposal to OERI, CREATE set as its mission the task of conducting research, development, and dissemination to assist U.S. school districts in the three ways noted in Table 1 (see page 8).

In the fall of 1990, CREATE was funded as a research and development center to address that mission for a period of five years and a total anticipated budget of $5.2 million. Thus, CREATE has completed two years of effort toward fulfilling its mission, is midway through its third year, and is planning for its fourth year of work.

Of its approximately $1 million annual budget, roughly $600,000 is available for direct R&D efforts. (Overhead, administrative costs, report development, etc., use up the remaining funds.) The net result is that CREATE has been able to support approximately 7-10 projects per year with annual budgets typically in the range of $50,000-$100,000 dollars. Routinely, the projects have worked with small professional staffs of 1 to 3 persons with some assistance from consultants. Typically, professional staff members devote in the range of 20 percent to 40 percent of their time on an individual development project, and no project employs a professional on a full-time basis.

As the budgets and staff commitments suggest, in any one year each project represents a relatively small investment. While the results of individual projects are expected to be important in their own right, this small investment suggests that for CREATE to succeed as a whole requires that the results cumulate across the years and that the individual projects serve a larger whole.

PROGRAMS

Recognizing that individual projects were and are limited in what they can produce, the initial proposal called for incorporating projects into programs. As originally configured, CREATE was comprised of four programs (see Table 2, page 9).

As Table 2 shows, the program format for CREATE has evolved slightly. Some program titles have changed a little, and the list order of the programs has changed as well (e.g., the original Program 3 has been renamed and shifted to the fifth program in the list); those changes are small, but have resulted in some confusion. The biggest change regards splitting the program Improvement of Evaluations of Administrators, Support Personnel, and Schools into two programs. This division of Program 2 into two separate programs (programs 2 and 3) was done because the school evaluation program is substantive in its own right and the separation provides greater visibility to school evaluation work. Thus, despite changes in appearance, CREATE's programmatic thrusts have been consistent.

When viewed from the present program format, the number of projects supported by year provides some perspective of emphases given to the respective programs. As Table 3 (page 10) shows, of the five programs the teacher evaluation program is CREATE's top priority. Table 3 also provides a quick perspective of CREATE's efforts to allocate funding in a way that enables CREATE to use its people and resources effectively across the five years to deal with a broad range of issues. For example, the start-up of efforts in the administrator and support personnel program coincides with reduction in efforts in special projects and teacher evaluation.

The fifth program listed in Table 3, National Evaluation Resource Services, is CREATE's program to disseminate information to CREATE's national constituency. As important as it is to the overall success of CREATE, that program does not have an R&D focus. For that reason, the program will not be addressed further in my comments here.

OVERVIEW OF CREATE PROJECTS

CREATE contains a total of 15 projects: 4 have reached completion, 9 are currently in progress, and 2 have been proposed for start-up in Year 4. Table 4 (page 11) lists these projects, organized by program.

Model Building. CREATE's original proposal to OERI argued that current evaluation practices are inadequate for the needs for personnel and school evaluations. Because evaluation models serve as the conceptual frameworks for conduct of evaluations, it is not surprising that CREATE's original proposal set as a main objective to carefully evaluate existing models and strategies and either construct new models or significantly improve existing ones.

This model-building process has four distinct strands that individually address teachers, administrators, support personnel, and schools. A clear expectation of CREATE is the production of new or improved models for evaluation for each of those four targets.

Table 5 (page 12) depicts the model development efforts as originally proposed. As that table shows, six separate projects were explicitly slated for model development work. As CREATE has evolved, three projects in addition to those noted in the table have been added to the dimensions of this model building process; two of the three are in teacher evaluation, with the third in school evaluation.

1. The Grounded Theory of Teacher Evaluation project (a two-year effort), through analysis of actual teacher evaluation systems, is developing an evaluation framework, including process and instruments, that school systems can use to evaluate the products of their school's teacher evaluation system, identify problem areas, and move toward improvement of an existing system.

2. The teacher self-evaluation project is developing a conceptual framework and compiling strategies to assist teachers in assessing their own progress and effecting change in their instruction.

3. Within the Improvement of School Evaluation program, the School Report Card project is developing a framework for schools to use in reporting results of their school evaluation efforts. This model, important in its own right, will also provide substantial information for development of the school evaluation model and, ultimately, is likely to be embedded within CREATE's school evaluation model(s).

As depicted in Table 6 (page 13), these three additions provide a much more comprehensive model structure for CREATE.

Model Development Methods. Because a great deal of CREATE's energies have been oriented toward the development of evaluation models, I think it is useful to try to depict, in a rough sort of fashion, the general approaches employed. Figure 1 (page 19) depicts the two basic strategies

BEST COPY AVAILABLE

employed. One strategy, the one on the right, begins with review and analysis of extant models. A primary product of that review is a consumer report on the models reviewed. That consumer report, along with feedback from practitioners and researchers in the field, is then used as a basis for development of a model or set of models. Dr. Scriven's Teacher Evaluation Models Projects is one example of this strategy.

TEMP, as it is called, has reviewed 15 models or approaches for teacher evaluation and developed a consumer report, which was distributed in preliminary form as TEMP A. That consumer report in expanded form is presently being prepared as a book, *Teacher Evaluation: A Consumer's Guide*, for publication by Corwin Press/Sage. Information from that report, along with input from individuals who have read the report, will be used to frame one or more new or revised models for teacher evaluation during Year 4 of CREATE. Those models will undergo field testing and revision during the fourth and fifth years of CREATE.

That same pattern, with some modification, is employed in the development of school evaluation models. This effort is characterized by the development of 2 separate but related consumer reports. First, during Years 1 and 2 of CREATE, the School Evaluation Models project analyzed 8 regional accreditation association models, 22 state models, and 15 local evaluation models for evaluating schools, a total of 45 separate models. The analysis of these models has been reported in *Consumer Report on School Evaluation Models*, which is currently undergoing review for publication by Corwin Press.

Second, as previously noted, beginning in Year 2 and continuing in Year 3 of CREATE, a consumer report on school report cards is being prepared. That report, in addition to being distributed as a separate entity, will be coupled with the consumer report on school evaluation models as input for the development of improved model(s) for teacher evaluation.

The left-hand strand of Figure 1 depicts the general strategy employed by the other four model development efforts. Those efforts have progressed less far to date, but their general patterns for model development are clear. As a beginning point the models using this strategy are engaging in a review of literature and a gathering of information about extant practices. Thus, for these projects the primary initial products are literature reviews and status reports. As is the case with the projects that have begun by conducting a review of models, the status reports and initial model plans are subjected to intensive review by evaluation experts and practitioners in the schools.

In the case of the teacher self-evaluation project, teacher self-evaluation models do not exist as distinct entities. Rather, numerous strategies for teacher self-assessment are described in the literature. To capture both the strategies promoted in literature and the additional strategies employed by teachers, the project is following up its literature review with teacher focus group interviews. The resulting combination of literature and field-based input is expected to provide a sound basis for both explicating most promising practices and setting forth a sound model.

Similarly, for administrator evaluation there does not exist a distinct set of frequently used administrator models. Thus, that project is reviewing literature along with direct analysis of models and practices employed in a variety of school settings.

In the case of support personnel, Drs. Stronge and Helm joined CREATE with a model for evaluation of support personnel in hand. Thus, their efforts focus on enhancement of this model. They are conducting a careful review of literature, along with thorough analysis of the extant model, to identify additional ideas and materials that can be used to improve the existing model.

The Grounded Theory of Teacher Evaluation project is probably best described as a hybrid of the two methods I described. This project began based on a clear knowledge of the literature base and current models for teacher evaluation. It has been gathering data from currently existing teacher evaluation systems in the local school systems. These practices, including instruments, procedures, support materials, and products of the systems as they actually occur, are being carefully scrutinized and organized, following grounded theory research processes, into input, process, and output categories. Outputs in turn are examined in light of *The Personnel Evaluation Standards* as a beginning point for analysis of the system. The analysis model is now emerging through this process.

Beyond Model Building. Four projects are not engaged in this model development effort. Rather, they are focused directly on gathering information, implementation of a model, or development and application of materials that are likely to be of assistance in special circumstance evaluations. The products of those projects provide insights and techniques that are likely to be quite useful to the other model development projects.

One project, Expert Science Teacher Evaluation (Years 1-3), has focused on implementation of evaluation procedures within a curricular area, science, and from a particular model for teacher instruction, the constructivist perspective. That project is engaged in moving teacher evaluation efforts forward in two strong ways. First, it promises to provide substantial information about characteristics of expert science teachers and how they differ from novice teachers. Second, the instruments and processes can be used directly by teachers as formative evaluation tools to help improve classroom instruction (in essence, teacher self-evaluation tools).

A second project, Databases of Practices in the Evaluation of Educators and Schools, engaged in acquisition of information to be used by the other projects. It surveyed schools nationally and elicited information from numerous reports to provide an overview of educational climate information about each state. Additionally, as part of CREATE's initial attempts to focus on development of relations with 10 states per year, schools in 10 states were surveyed to provide background information about evaluation practices in the respective states. Information from both survey efforts, along with substantial additional materials (e.g., a cross project glossary of evaluation terms), is stored in databases for use by CREATE researchers.

A third project developed guidelines for the evaluation of teachers who teach at-risk students. That project identified characteristics deemed to be important in the instruction of at-risk students and used a broad cross section of teachers, administrators, researchers, and others to help determine the appropriateness of the criteria. The result is a draft set of guidelines that can be used in a variety of ways in selecting, evaluating, and assisting teachers and administrators as they work with at-risk students.

A fourth project, Adapting W. Edwards Deming's Statistical Process Control Model, actively tested application of the Deming model for Total Quality Management (TQM) in school settings with teachers and principals. This TQM model has been widely acclaimed in business and industry and has been advocated for use in public schools. Like the Expert Science Teacher project, the primary products of this project were instruments and materials and a handbook for use in implementing that approach in the schools.

CREATE Products. Though the completed, field-tested models are primary objectives of CREATE, each project is producing other products that are important in their own right. Though CREATE is still young, numerous papers, instruments, and other materials have been produced. These products have sparked interest, debate, and a rethinking of the bases for evaluating teachers and other school personnel.

Table 7 (page 14) provides a summary of principal interim products being produced in the development efforts. This list shows extensive productivity in identification, explication, and evaluation of current evaluation models in each of the target areas. Additionally, the list demonstrates the wide array of information and materials that flow out of the projects. For example, a person interested in teacher evaluation can, through these projects, learn the terminology, access current literature, identify major duties of the teacher that should be included in teacher evaluation, learn the relative strengths and weaknesses of current teacher evaluation models, conduct an analysis of an existing teacher evaluation system, access instruments for use in teacher evaluation, learn effective self-assessment strategies to improve instruction, and a lot more.

AN OVERARCHING ISSUE

Most of the work completed at this point in time can be called foundational. One important result of this effort has been confirmation of assumptions stated in CREATE's initial proposal. There it was argued that personnel and school evaluation models and practices are in need of substantial improvement. At least four findings common across the model development projects, provided in Table 8 (page 17), confirm those positions for virtually every target group being addressed by our studies.

The respective projects note such things as the misuse of instruments (e.g., the same instrument used for all teachers, and even counselors); restricted methods for data collection (e.g., teachers observed by principals, principals observed by superintendents as the only input for the evaluation); and a poor match between job description and data collected, with too little attention to the varied aspects of the teacher's, administrator's, or counselor's work (and correspondingly too much attention to factors that research says correlate with job success).

There is an old adage that says "if it ain't broke, don't fix it." Anyone reading that litany of ills would, you would think, quickly conclude that the evaluation systems are broke and we must fix them. However, each project has noted instances, models, and practices where things have been done correctly. More generally, results from the databases project's survey of schools in ten states indicate that representatives of the schools do not view their local evaluation systems as "being broke." These representatives depict evaluations as occurring in their schools in much the same way the development projects describe, but they arrive at a markedly different conclusion.

Findings from the survey suggest that school representatives are aware of the problems with evaluation, but believe that the evaluation system in their own school works. That perception may be a substantial barrier to changing existing practices in schools, particularly since administrators are likely to be the gatekeepers of change in evaluation processes. Certainly, the findings should make us aware of the need to work carefully with the schools in moving to change the current systems.

Table 1

CREATE's Mission

Conducting research, development, and dissemination to assist U.S. school districts to

1.	validly and reliable evaluate	→ teacher → administrator → institutional performance
2.	effectively use the evaluation results to improve educational services to	→ students → communities
3.	credibly and systematically assure accountability to	→ constituents → sponsors

Table 2

CREATE's Programs

Original Configuration	
1	Improvement of Teacher Performance Evaluations
2.	Improvement of Evaluations of Administrators, Support Personnel and Schools
3.	Products and Services/Dissemination
4.	Theory Development and Special Projects
Current Configuration	
1.	Improvement of Teacher Performance Evaluations
2.	Improvement of Administrator and Support Personnel Evaluations
3.	Improvement of School Evaluations
4.	Theory Development and Special Projects
5.	National Evaluation Resource Services

Table 3

CREATE's Programs

Program	Project Years	Number of Projects				
		Yr 1	Yr 2	Yr 3	Yr 4	Yr 5
Teacher Evaluation	12	2	3	4	2	1
Administrator and Support Personnel Evaluation	6			2	2	2
School Evaluation	5	1	1	1	1	1
Cross-Cutting Theory and Special Projects	9	3	3	1	1	1
National Resource Services Center	5	1	1	1	1	1

Table 4

CREATE Projects by Program

Teacher Evaluation	**Administrator and Support Personnel Evaluation**	**School Evaluation**	**Cross-Cutting Theory and Special Projects**	**National Resource Services Center**
Teacher Evaluation Models Project (Years 1-3)	**Administrator Evaluation (Years 3-5)**	*School Model (Years 1-2)*	*Databases of Practice (Years 1-2)*	**Dissemination of Research and Products (Years 1-5)**
[Improved Teacher Evaluation Models Project (Years 4-5)]	**Support Personnel (Years 3-5)**	**School Report Card (Years 2.5-3)**	*Deming's Stat. Control Model (Years 1-2)*	
Grounded Theory of Teacher Evaluation (Years 2-3)		[School Model Development (Years 4-5)]	*At-Risk (Years 1-2)*	
Teacher Self-Evaluation (Years 2.5-4)			**Cross-Cutting Theory (Years 3-5)**	
Expert Science Teacher Evaluation (Years 1-3)				

Note. Projects in **bold** are currently in progress, those in *italics* are completed, and those [bracketed] are included in current plans for the future. One project, **Evaluations of Nationally Significant Evaluation Programs**, is not included in the list. During the first two years of CREATE it was conducted under the auspices of CREATE but funded separately.

BEST COPY AVAILABLE

Table 5

CREATE's Proposed Model Development Efforts

Evaluation Model	**Projects**	**Project Years**
Teachers	Teacher Evaluation Models Project Improved Teacher Evaluation Models Project	5
Administrators	Models for Administrator Evaluation	3
Support Personnel	Model for Evaluations of Professional Support Personnel	3
School	Models for School Evaluation	4

BEST COPY AVAILABLE

Table 6

CREATE Model Development Efforts

Evaluation Model	Projects	Project Years
Teachers	Teacher Evaluation Models Project Improved Teacher Evaluation Models Project	5
Teacher Evaluation Systems	Development of a Grounded Theory of Teacher Evaluation	2
Teacher Self-Evaluation	Development of Classroom Assessment Techniques for Teacher Self-Evaluation	2
Administrators	Models for Administrator Evaluation	3
Support Personnel	Model for Evaluation of Professional Support Personnel	3
School	Models for School Evaluation Improved and Tested School Evaluation Models	5
School Report Cards	Development of a Research-Based School Report Card	

BEST COPY AVAILABLE

Table 7
Principal Interim Products and Working Documents
in CREATE's Model Development Efforts

[Titles in bold are currently available; non-bold titles are either in process or under review.]

I. Teacher Evaluation Program

- A. Teacher Evaluation Models Project
    1. ***Teacher Evaluation Glossary***
    2. **"Duties of the Teacher"**
    3. **"Building Teacher Evaluation Systems: Finding the Right Foundation"**
    4. *Teacher Evaluation Bibliography and Abstracts*
    5. Temp Memos **A,B,**C,D,E,F,G,H
    6. *Teacher Evaluation: A Consumer's Guide*
    7. Teacher Evaluation Assistance System Handbook

- B. Development of a Grounded Theory of Teacher Evaluation
    1. "Evaluation For Effective Teaching"
    2. **"Competing rationales for evaluating teacher performance"**
    3. **"Educational personnel evaluation"**
    4. **Format for Documenting a School or District Teacher Evaluation System (Open coding instrument)**
    5. **Format for Documenting a School or District Teacher Evaluation System (Axial coding instrument)**
    6. **"What fits under the label of teacher evaluation?"**
    7. Coding instruments to analyze teacher evaluation systems
    8. Fast Response Survey instrument for a national survey of elementary teachers (Conducted by Westat for NCES)

- C. Development of Classroom Assessment Techniques for Teacher Self-Evaluation
    1. **"Self-assessment in narrative/qualitative studies"**
    2. **"A model of teacher self-assessment"**
    3. **"Beginning teachers and self-assessment"**
    4. "Formal self-assessment practices"
    5. Literature Review on Classroom Assessment Techniques
    6. Status Report on Teacher Self-Assessment Techniques
    7. Handbook of Teacher Self-Assessment Strategies

- D. Expert Science Teacher Project
    1. **Science Classroom Observation Record (SCOR)**
    2. **Student Outcome Assessment Rubric (SOAR)**
    3. Expert Science Teacher Evaluation Model (ESTEM)
    4. "The Process of Developing a Composite of Expert Science Teaching"

5. "Toward a description of excellent science teaching: Appropriate and productive analyses of qualitative data"
6. "Perspectives on methods and procedures for evaluating teachers at different stages of professional growth"

II. Administrator and Support Personnel

A. Administrator Evaluation
  1. **"Principal evaluation: New directions for improvement"**
  2. State of the Art of Principal Evaluation
  3. State of the Art of Superintendent Evaluation
  4. Bibliography on administrator, principal, and superintendent evaluation
  5. A new model for principal evaluation
  6. A new model for administrator evaluation
  7. Manual for implementing the new model for principal evaluation
  8. Manual for implementing the new model for superintendent evaluation
  9. Publication on a field-tested model for superintendent evaluations
  10. Report of efforts to disseminate the new evaluation models

B. Support Personnel Evaluation
  1. **"A performance evaluation system for professional support personnel"**
  2. Survey Form for Evaluation of Professional Support Personnel
  3. Evaluator's Technical Manual
  4. Practical Guidelines in Implementing the Professional Support Personnel Evaluation Model (together with evaluation forms)
  5. Multimedia materials
    a. Slides depicting conceptual framework and practical applications
    b. Transparencies describing recommended evaluation forms
    c. Video materials depicting simulated evaluation settings
  6. "Performance evaluation for school counselors"
  7. "Special education support personnel evaluation"
  8. *Evaluation Practices for Professional Support: Bridging Theory and Practice*

III. School Evaluation

A. School Evaluation Models
  1. **Consumer Report on School Evaluation Models**

B. 1. School Report Cards
  2. Consumer Report on School Report Cards

C. 1. Improved School Evaluation
  2. Models for School Level Evaluation

IV. General and Special Projects

- A. Databases of Practices in the Evaluation of Educators and Schools (Database Project)
  1. **Glossary of Terms: CREATE**
  2. **Survey Instrument for States Database**
  3. **States Database**
     - a. **Print version**
     - b. **Hypercard version**
  4. **Report of the 10 State Survey - Abbreviated Overview**
  5. **Summary of Reported Practices in Teacher Evaluation**
  6. **Summary of Reported Practices in School District Evaluation**
  7. **Summary of Reported Practices in Administrator Evaluation**

- B. *Adapting W. Edwards Deming's Statistical Process Control Model: Trainer/Team Member Manual*

- C. ***Working With At-Risk Students and Families: Guidelines for Evaluating Teachers, Principals, Counselors, Psychologists, and Social Workers***

Table 8

Factors that suggest school evaluation processes are neither valid nor conceptually sound.

I. Personnel evaluations are not tailored to fit the particular position being evaluated.

II. The evaluations do not provide feedback for improving performance.

III. Information obtained in evaluations is not carefully analyzed.

IV. Those conducting evaluations are not trained in how to conduct evaluations or analyze evaluation data.

Table 9
Findings From CREATE's Survey of Schools in Ten States

Survey Findings Consistent with CREATE Development Project Findings

1. 98 percent of the school districts conduct teacher evaluations, 96 percent conduct administrator evaluations, and 65 percent conduct school evaluations.

2. Classroom observation alone or in combination with management by objective (MBO) information and interview is the dominant process for teacher evaluation. Teacher self-evaluation is used in about a quarter of the schools, and student learning outcomes are used in slightly more than 10 percent of the cases. Use of peer ratings and parent/teacher ratings is almost negligible (less than 5 percent).

3. Principal observation of teachers and superintendent visits with principals are the predominant methods of evaluation.

4. They acknowledge possible problems "such as lack of comprehensiveness and potential bias in the model, use of irrelevant evaluation criteria, and . . . difficulty of administration . . ."

Survey Findings That are Contrary to Development Project Findings

Most respondents

1. identify professional improvement, instruction for teachers, and performance for administrators as the purpose of the evaluation

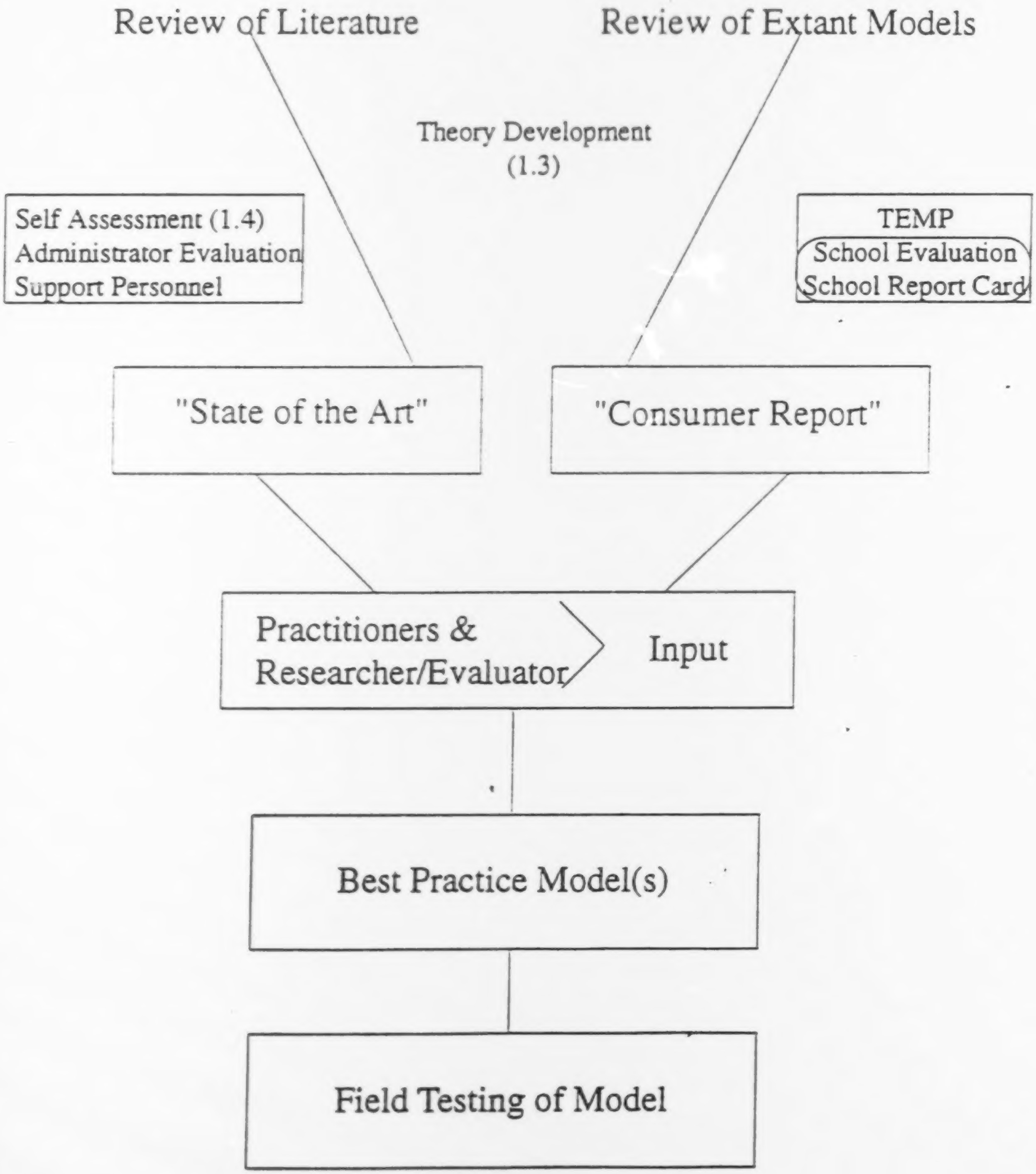
2. consider their teacher evaluation systems reliable and valid

3. did not find their administrator evaluation process to be bias prone, unreliable, or low in credibility. Neither did they consider the process to be unpopular.

While admitting to potential problems in the evaluation system, these problems were not seen as weaknesses their own systems.

BEST COPY AVAILABLE

Figure 1
Evaluation Model Development Methods


Review of Literature
Review of Extant Models
Theory Development
(1.3)
Self Assessment (1.4)
Administrator Evaluation
Support Personnel
TEMP
School Evaluation
School Report Card
"State of the Art"
"Consumer Report"
Practitioners &
Researcher/Evaluator
Input
Best Practice Model(s)
Field Testing of Model

ATTACHMENT 1

THUMBNAIL SKETCHES OF CURRENT DEVELOPMENT EFFORTS

1. Teacher Evaluation Models Projects (life span of 5 years). The first project (TEMP) will produce an analysis of current teacher evaluation models, identify strengths and weaknesses, and select the best models for use in preparing improved teacher evaluation models. The second will carry out the task of developing actual models for conducting teacher evaluation.

2. Teacher Self-Assessment Project (life span of 2 years). This project begins with a literature review, identification of teacher self-evaluation strategies, and conceptualization of a self-assessment model. Input from teachers will supplement lit review, but the outcome is expected to be a model for self-evaluation practices including a handbook of self-evaluation practices that teachers can employ.

3. Expert Science Teacher Project (2+ years). This project proceeds from an existing model on teaching practice (constructivist model) and has actively worked to determine the characteristics that distinguish expert teachers from novice teachers. Classroom observation instruments have been created to assess these differences and are expected to be coupled with overarching strategy to provide an expert teaching evaluation model that can be employed to facilitate professional growth and illustrate excellence. (Though supported initially through CREATE, most of the current work is being conducted under the auspices of the Dwight D. Eisenhower Fund for mathematics and science.)

4. Evaluation Theory Development Project (2 years). This project began with a study of extant teacher evaluation systems as they exist and are reported to be conducted in school settings. A grounded theory approach is being used in the conduct of this analysis and has resulted in first describing constituent characteristics of these systems, analyzing those features, and selecting the most salient features of the systems to provide input, process, and outcome characteristics of the systems. These characteristics are being joined, logically and empirically, to standards for the evaluation of personnel. The result of this project will be an evaluation model for use by schools. This model will include instruments and procedures that schools can use to evaluate and improve their existing teacher evaluation systems. The model calls for the school to review the outcomes of its teacher evaluation system and determine whether the system is performing satisfactorily to meet the standards. If the system fails to fully satisfy all the standards, the school can use instruments and procedures provided as a part of the model to trace back through the system and identify processes and inputs that are responsible for unsatisfactory performance of the system. Resource information and materials will be provided as part of the model to assist the schools in making necessary changes to improve its working model. As such, this project will develop a model for evaluating teacher evaluation systems rather than a model for conducting teacher evaluations.

5. Administrator Evaluation Models Project (3 years). This project will review the literature on administrator evaluation, available administrator evaluation models, and current practices to first provide a "state of the art" perspective on administrator evaluation. This state of the art will then be used to shape a working model for administrator evaluation.

6. Professional Support Personnel Evaluation Model Project (3 years). This project begins with a preliminary model for evaluation in hand. The project is reviewing literature on support personnel evaluation and conducting an intensive analysis of the existing model. That information together with input from the field will be used to improve the existing model.

7. School Evaluation Models Projects (5 years). This set of three projects is intended to culminate in one or more strong models for school evaluation. The first project evaluated existing models for school evaluations and culminated in a consumer report. The second project is currently evaluating school report cards (school report cards report school evaluation information) and will culminate in a consumer report on school report cards. Input from both projects will be used together, with practitioner and researcher input, to create an improved model or models for school evaluation.

APPENDIX C

THE FOUNDATIONS OF EDUCATIONAL ACCOUNTABILITY

Michael Scriven
Western Michigan University

THE FOUNDATIONS OF EDUCATIONAL ACCOUNTABILITY

Michael Scriven
Western Michigan University[1]

INTRODUCTION

This paper is directed towards identifying underlying theoretical assumptions of the many CREATE projects, partly with the aim of checking those assumptions for validity but also with the aim of producing some results of value to the projects by:

- *relating* them
- *contrasting* them
- *analyzing* them
- placing them into one or another kind of *perspective*.

From this process it is desirable that there should emerge:

- suggestions for *improvement*
- suggestions for *integration*
- *criticisms* of some approaches and *endorsements* of others
- a better basis for *reporting* out to our audiences and potential clients.

It may be just as well to start with some disclaimers. Writing a successful paper with the above mission strikes me as a formidable task, and three false starts on it have not increased my optimism. There is also a small matter of conflict of interest in that it seems clear that my own project is in a somewhat better position than the others with respect to the likelihood of severe criticisms emerging in the course of the present paper. The best solution to these difficulties which occurs to me is that we should regard this as a joint project. This paper should stimulate some thoughts which you can bring to future discussions, where I will try to improve it in the light of your reactions.

POSSIBLE APPROACHES

One might undertake this task simply by analyzing the methodological, logical, and/or philosophical presuppositions of the approaches of each project. I believe that would be quite fruitful, and in fact I have tried it. In the end, however, it leads to a fragmentary and mainly critical effort. I found it hard to bring the pieces together and the result exacerbated my feeling of unease about being the only one in a position to criticize the other projects.

The alternative I have chosen is to begin with a general definition of accountability, and see if that has any significance for the above tasks. One of the main concepts that comes into accountability is evaluation, so I next outline what seems to me to be the best overall general theory of evaluation, and apply that to our tasks—the missions for our projects—to see if that effort yields pay-off in terms of the goals of this paper. Both accounts—the general definition of accountability and the general theory of evaluation—are speculative and need improvement, but perhaps they will serve as a useful starting point for a discussion.

---

[1] Mailing address: POB69, Point Reyes, CA94956. Thanks to Geneva Haertel for valuable comments on an earlier draft.

ACCOUNTABILITY

CREATE is a center for research on educational accountability, yet we have rarely addressed that concept directly. We often hear it said that one of the functions of evaluation is to serve the need for accountability. This is true enough, but it rather suggests that accountability is an intuitively clear notion, which may not be so.

I suggest that **accountability** in the sense in which we now use it—one that goes slightly beyond the usual dictionary definitions—means, roughly speaking, **demonstrable and creditable responsibility**. The dictionaries stress the responsibility part, often using the term "answerable." But we normally mean more than this when we say that persons should be held accountable. We normally mean that a steward's or agent's performance should not only be one for which s/he is *answerable*, but that the answer is *satisfactory* and *that it can be shown to be satisfactory*. The term "satisfactory" here means "at least as good as could reasonably have been expected"—in other words, *creditable*. Of course, nothing can be shown to be creditable without some kind of evaluation, since "creditable" is an evaluative term.

Note that "creditable" does not mean, in this context, that the results of a highly accountable project were desirable or beneficial—although of course we hope for this—only that they were as good as they could reasonably have been expected to be. This relates to the next point.

The other half of "answerable," besides the part that requires the demonstration of merit—of at least the effort if not the outcome—relates to the matter of what questions have to be "answered." The dictionary suggests that these relate to the obligation to "report, explain, or justify."[2] Now *bare* reporting as in "I have spent the money for the project" is not what we have in mind when we talk about being accountable. Even *explaining* of certain kinds such as "I spent the money on gambling debts" is not judged to be a proof of accountability, but rather the reverse. The term "justify" comes closer to the key notion here. And of course, "justification" is an evaluative term, although it is not a term from the basic evaluative vocabulary of good and bad, better and worse, right and wrong.[3]

Hence, the fundamental concepts required to apply the concept of accountability are the notions of evaluation—in the sense of establishing creditability and justifiability—and demonstration, the latter being part of the generally accepted notion of the mission of serious personnel and program evaluation. Thus, the kind of evaluation referred to here is what might be called scientific or disciplined evaluation—by contrast with the pseudo-disciplines of wine-tasting or art criticism.

But you will have noticed that the *particular* kind of evaluation involved here goes somewhat beyond the standard types involved in program evaluation of the present age. There is an element of the ethical involved, an element of moral assessment of the exact kind with which the courts are frequently involved when matters of responsibility or culpability come before them.

---

[2] This quote is from the definition of "accountable" in the ***Random House Unabridged, Second Edition***.

[3] Reference should be made here to a serious work that takes a rather different approach: ***Accountability in Education: A Philosophical Inquiry*** by Robert B. Wagner (Routledge, 1989)

Hence, the first point suggested by this analysis is that:

1. Evaluators of service and training programs who claim to be looking at accountability should extend their skills to handle judgments of responsibility and culpability.

While we do see evaluations which undertake this dimension of the task, e.g., from GAO, it is not a universally accepted standard, partly because it requires the evaluator to take a moral stance in his or her own name. One often finds it absent from evaluations done by many centers and agencies working in the field of administrative or managerial or governmental policy analysis or staff training—areas where the notion of accountability is extremely important.

It is clearly true that if we adopt this posture, we will need to give our students and those whom we train on the job some leadership in exploring this matter. Not only the excellent legal materials in the area of liability, but also the excellent body of applied ethical materials that have been developed in the last couple of decades, make this a manageable task. It is in fact one for which we should be held increasingly accountable.

Accountability is ultimately *someone's* accountability. If programs are held 'accountable', what this means in practice is that their managers are being held accountable. It is generally thought by managers to be good management practice to pass part of their responsibility on to the rest of the staff, and not just the administrative staff but the certificated and other support staff. It follows that in program and institutional evaluation—in education or outside it—since we normally judge it appropriate that programs and institutions should be held to accountability standards, there must be some attention to the accountability of the personnel who make up the program. Now, to judge their accountability one must be in a position to judge, amongst other things, whether they are performing at a reasonable level of effort and competence. The basis for such judgments about personnel, as for programs, is not found in the usual program evaluation training manuals or courses.

The general version of this point of course implies an essential connection between program and personnel evaluation, a connection which has long been slighted.[4] That point suggests the possibility of some linkages between CREATE projects that may provide useful insights and perspectives, e.g., between the school evaluation project and the administrator evaluation project. I want to develop the foundations for a considerable further generalization of this kind of point and I'll take it in two stages.

The first stage, which we have already broached, begins with the need for bringing skills in ethical appraisal or moral judgment into program evaluation.[5] It is clear enough that we must extend the point to make clear that:

---

[4] This is not the only case of costly isolationism in the development of program evaluation. Here is a reference to it and to another example. "To mention two of many possible examples: product evaluation was not even considered as an appropriate and illuminating exemplar *for* the methodology of program evaluation; and personnel evaluation was not seen as an essential element *in* program evaluation." ***Evaluation Thesaurus 4e***, (Sage, 1991).

[5] Ernie House was an early advocate of this, for example in ***Evaluating With Validity***, Sage, 1980.

2. At least in some loose sense, personnel evaluation must involve ethical appraisal.

We have often sensed this, but it has not often been spelled out, particularly for formative evaluation.[6] When we talk of holding teachers accountable, e.g., for children's learning, we immediately run into the question of *how much* of that learning they can be held accountable for, given the many other factors at work. It is clear that this is an argument about what constitutes a *fair* approach to teacher evaluation, that is, about the ethical dimension in teacher evaluation. Again, in teacher evaluation, we should probably be considering the question whether teachers should treat moral education as an obligatory part of the across-the-curriculum repertoire.

Ethical appraisal is a far cry from the concerns of most personnel psychologists and many educational personnel evaluators, seems to be an essential part of our task in dealing with diverse problems in educational accountability. Are we well equipped for this in our work at the Center? Should we be specifically addressing this dimension in all our efforts, since it clearly comes into program and institutional evaluation as well as personnel evaluation?

The second stage requires us to get into general evaluation theory, but it is useful if we first remind ourselves of the overall configuration of the efforts at the Center.

THE ACTIVITIES OF THE CENTER

The projects at this R&D center involve three types of evaluation, as one conventionally distinguishes these things. First, there are a number of exercises in the area of personnel evaluation—namely, the evaluation of teachers (including teacher self-assessment), administrators, and support personnel. Second, there is some institutional evaluation—we have two projects concerned with the evaluation of schools using report cards, indicators, and other models. Institutional evaluation has not been generally recognized as an independent type of evaluation, and is normally taken to be close to or a version of program evaluation. That may be an assumption we need to question; after all, institutions usually have buildings and when did you last run through a buildings checklist when doing a program evaluation?

3. Institutional evaluation (e.g., of schools) is more than an application of program evaluation.

Third, there is some work on evaluation theory, of which this effort is a small part: if you think it is turning up some useful results you may feel that that in itself is an interesting finding:

---

[6] The ***Standards for Evaluating Personnel*** include Propriety which covers some ethical considerations but not much detail about what it's reasonable to require in specific matters such as off-campus public life, a frequent source of contention and resentment. Of course, teachers are, in cases that seem extreme to the administrators involved, fired for 'moral turpitude' but the term is rarely defined in contracts. (There is a good body of case law which gives the accepted instances but no general definition.) But we are here going much further, arguing for the introduction of moral assessment into appraisal for professional development, not just as grounds for dismissal. (We have certainly not done a thorough job in this area, in the Duties of the Teacher document, the key to one of our recommended approaches.) A typical and troubling problem that would have to be faced is the question of whether all teachers have an obligation to provide moral education of some kind, as an across-the-curriculum subject, as private denominational schools have often argued. Recent efforts in this direction have been along the politically correct but dubiously valuable and very expensive path of required courses in multi-cultural awareness.

4. Work on evaluation theory may contribute substantially towards the improvement of evaluation practice.

And finally, there are some infrastructure efforts, most notably the administration and dissemination without which we could not get anything done or implemented.

One might take this moment of overview to comment on some aspects of the mission suggested by our name that are missing from this set of projects, through limitations of resources and turf distinctions that have some but perhaps not an adequate justification. Educational accountability, speaking comprehensively, would also involve some other components. (These are not put forward as things we can afford to do with our present budget, but as matters that should be addressed in any further funding of accountability in education.)

5. A complete approach to educational accountability would involve curriculum evaluation.

Curriculum evaluation is an important part of education which is supposed to be accountable to the society, in its efforts to meet national and global needs. This is particularly evident since the present situation in curriculum evaluation is desperately and expensively lacking in any systematic approach.[7]

6. A complete approach to educational accountability would involve some evaluation of educational policy at various government levels.

The reason for this is that there can hardly be any accountability without a policy of requiring and supporting it. A simple example concerns the usual practice in post-secondary policy of using enrollment figures as a needs assessment. The result is of course that departments who set high standards get low enrollments and can go out of existence. In the extreme, which is common, the central administration's policy undercuts any attempt at demonstrable, creditable, responsibility for academic standards.

7. A complete approach to educational accountability would involve extending personnel and institutional evaluation to include those ghosts at the banquet of all other such efforts, the evaluation of school boards, state education agencies, and the federal officers in charge of education and education-related departments[8].

THE FOUNDATIONS OF EVALUATION

The discussion of accountability has suggested some directions to go if we are to talk of integrating or even relating our projects; or of extending them into territory that they should explore but which, following conventional wisdom, they have not so far investigated to any substantial extent. However, the most fundamental suggestions for connections and extensions emerge from a certain view of evaluation itself. While some of you are familiar with other elaborations of this "transdisciplinary view" of evaluation, others are not and since I have continued to develop it beyond the versions available in print, in particular for this occasion, perhaps it is not inappropriate to provide a brief review.

---

[7] The recent efforts, e.g., in science and math, have been worthy proposals but not based on any discussion of the appropriateness of the present division of the turf, since they never raise seriously such questions as whether there is already far too much curriculum time spent on science and math. (Indeed, they implicitly assume that there really needs to be more time for those activities.) See the forthcoming Expert Panel report on federal efforts in the areas of science/mathematics/engineering/technology education.

[8] There are, for example, 12 federal agencies making major investments in the sub-area of education which covers science, math, engineering and technology education. ("Major" here means $24bn per annum, according to the Expert Panel review just completed.)

The transdisciplinary view is in essence not too complex; it is an "Emperor's new clothes" kind of theory. Of course, that was also true of the special theory of relativity, so it isn't necessarily the same as triviality. I'll try to compress the elements of the transdisciplinary view (TDV) into eleven short points, which I hope will strike you as plausible, since I then propose to derive several consequences for our activities at the Center.

I. Evaluation is an academic discipline that meets the appropriate disciplinary standards of objectivity and comprehensiveness. If it couldn't do that, no other disciplines could do so, since each of them depends absolutely on the internal use of evaluation for its own validity (the community of scholars making up the discipline evaluates proposed findings through peer discussion); and each discipline depends absolutely on evaluation for passing along its findings (the evaluation of papers submitted for publication, the evaluation of candidates for positions in research centers). This use of evaluation is 'intradisciplinary evaluation' of the academic subspecies[9].

II. Evaluation is also a practical discipline, long exhibited in the practices of technology and performance improvement. Its validity in that role is illustrated by the fact that every one of those who are publicly sceptical about the possibility of objectivity in evaluation act in private in a way that makes clear they believe in the objectivity of many examples of evaluation, as they make clear from reading and heeding the reports in ***Consumer Reports***, ***Road & Track***, or the media reporting on FDA or NIH evaluations of drugs or surgical procedures. Other branches of intradisciplinary practical evaluation are found in the crafts, the physical disciplines, and the "practical arts" of human affairs, the latter including include management, teaching, business operations, street smarts, planning, purchasing, cooking, policing, and so on.

III. Product evaluation, referred to in the last paragraph, is just one of several well-developed autonomous fields of evaluation, many of them making much use of tools from the social sciences. These include personnel evaluation, program, policy, proposal and performance evaluation, which, together with product evaluation make up what is conveniently referred to as the "Big Six." These are not in general more difficult than product evaluation; the evaluation of athletic performances, for example, might seem to be somewhat easier, but the formulas used by the IAAA to compensate for altitude and track condition are quite complex. Other autonomous fields include technology assessment, quality control, and sensory evaluation (e.g., wine-tasting). So the matrix of evaluation applications runs vertically through intradisciplinary evaluation, and horizontally across all of the usual disciplines and also across many autonomous applied evaluation fields which are typically interdisciplinary.

---

[9] Another way to put the point follows. The distinction between scientific (or other) disciplines and pseudo-disciplines, e.g., between astronomy and astrology, is based entirely on evaluation (of the quality of data, hypotheses, instruments, inferences, and theories), and the distinction between poor work within a discipline and the best work is itself evaluative. So the disciplines themselves depend for their existence and functioning as disciplines on the possibility of objective intradisciplinary evaluation.

IV. The unique subject matter of evaluation is not some segregated horizontal level or vertical slice of the phenomena of the social or physical or biological world, but simply one type of property that most entities in the world possess in relation to certain functions, needs, and contexts, essentially their merit or worth, their quality or value. This attention to one family of properties rather than to the phenomena of some separable slice of the world's events, makes evaluation akin to certain other disciplines, notably measurement (which is concerned with quantifiable properties), statistics (which deals with the global properties of countable ensembles), and logic (which is concerned with the validity of arguments, and thus is in a sense a field of applied evaluation). These fields apply across other fields—are used as tools by other disciplines—and are hence referred to as *transdisciplines*, by contrast with *interdisciplines* which focus on an area which is overlapped by more than one discipline, and *multidisciplines*, which use the methods of several disciplines. The validity of intradisciplinary evaluation (and logic) is absolutely required for the validity of all disciplines; other transdisciplines are not quite so universally relevant, but many traditional disciplines do depend on the validity of measurement and others depend on the validity of statistics.

V. Apart from the intradisciplinary use of the transdisciplines, and their and task-oriented use in the autonomous fields, each is also a discipline in its own right. In that role it devotes itself to identifying, developing, and investigating its own presuppositions, theories, and methods, and to delineating its fields of proper application (and its misapplication). That part of its effort is referred to here as 'the core discipline'. While skill in intradisciplinary evaluation (or logic) is learnt as part of the (mainly tacit) process of learning each discipline, an understanding of the transdisciplines themselves comes only from study of the core discipline as well as its many applications in the applied fields. This theoretical understanding is valuable not only in the service of greater overall understanding but because it leads quickly to powerful improvements in the applied fields. In the case of evaluation, these payoffs are in intradisciplinary fields of evaluation (e.g., improvements in the quality of evaluation of proposals, software, and completed research within the sciences) and in the autonomous applied fields of evaluation such as program and personnel evaluation (examples are given below); and also—in a special case of an intradisciplinary approach—to the discipline of evaluation itself, which leads us to develop standards and skills in meta-evaluation.

VI. The autonomous sub-fields of the transdisciplines—by contrast with their intradisciplinary applications—develop in two different ways. They sometimes emerge from practical necessity without benefit of the core discipline (as product evaluation emerged from product making); and sometimes—as in the case of biostatistics and quantum mechanics—their emergence is facilitated by and comes after considerable sophistication has been developed in the core discipline (in this case, statistics). Another example of emergence from practical affairs without assistance from a core discipline is provided by the emergence of logic from the practice of reasoning and argumentation. Product evaluation not only emerged from and as part of primitive technology but reached the status of a career many centuries ago with the Japanese sword-testers. Personnel evaluation, while it may have been a career as early as the Pharaohs, certainly became one in mid-century when several doctoral programs in psychology offered it as a specialty within industrial/organizational psychology.

VII. These "self-starting" applied fields like product or personnel evaluation can solve many practical problems, and they can avoid the major problem of getting bogged down in academic theorizing that is irrelevant to practical considerations. But this kind of development runs two risks, historically speaking: (i) the risk of laboring to reinvent a wheel discovered long ago in another applied field of evaluation; and (ii) the risk of developing a fundamentally unsound practice because of the lack of theoretical tools to do deep analysis of its underlying logic and assumptions. Thus even professional gamblers in the pre-probability days had a number of beliefs about sound strategies that they took to be well-based in experience, which were in fact unsound. Similarly, Consumers Union, which eschews discussions of methodology, makes some quite serious mistakes as a result[10] and we find similar errors in personnel evaluation, program, proposal and performance evaluation (some examples follow).

VIII. Evaluation's principal source of empirically-based value premises is needs assessment, although many of its value premises come from purely normative sources such as legal, ethical, and professional codes. This pervasive use of needs assessments is sometimes taken to be—but is in fact not—a sign that its findings are inescapably subjective. No-one has ever suggested that the same dependence of human medicine on needs assessment is a ground for dismissing its findings as inescapably subjective. Needs assessments, properly done, are a scientific investigation, part of evaluation as a discipline. We know how the body's need for vitamin C was determined, and needs assessments in the educational or social services areas should be determined in a similar way.[11] Moreover, evaluation can easily be applied to the needs of other animals, in the way that—in the analogy—leads to veterinary medicine; or to the needs of plants and planets, in the way that generates horticulture or ecobiology; or for that matter to the needs of cars or manufacturing processes, as studied in automobile maintenance research or industrial engineering.

IX. The fact that a field is clearly an evaluative field—in whole or part—does not lend the slightest credence to the conclusion that it meets any standards of quality at all. The extreme example is art and music criticism which is largely a parade of subjective preferences dressed up as evaluation (as history shows); wine-tasting as usually practiced is indistinguishable from wine-recognition (as the blind tastings show); literary criticism, apart from its major nonevaluative role as a process of enlightenment (instruction, information, education) is also highly evaluative, and in that role almost entirely lacking in objectivity (as the success of deconstructionism shows).

X. Conversely, however, the fact that such poor examples of evaluative subjects exist lends no credence to the idea that evaluation is essentially subjective, or else every grade given by every science instructor would be arbitrary, along with every interpretation of every phenomenon to which scientists have turned their attention. Constructivists who think that to be in fact the case have not yet seen that their position is self-refuting, along with and for the same reason as its many defunct predecessors in the history of epistemological scepticism. For one cannot advocate the view that all claims to truth are merely an expression of the way that the claimant has constructed reality—and hence no more valid than the apparently incompatible claims of another—without being open to rebuttal by turning the remark on itself. When we do so, it is clear that this claim entails that it is itself no more worthy of being taken seriously than its denial. But of course constructivists think their thesis expresses a profound truth, *far* more worthy of belief than the opposite view. Hence their position is self-contradictory, in that its acceptance entails its rejection.

---

[10] This issue is an aside, but for those interested, the most basic flaws are: (i) the low quality of its needs assessment, which comes nowhere near to meeting minimal standards for best practice needs assessments in program evaluation, a flaw which automatically corrupts its selection of products for evaluation; (ii) the increasing conflict of interest problems associated with both its huge book-publishing operation, which it fails to have externally and globally evaluated, the 'loose cannon' approach of the regional offices on local issues, and its massive adoption of many of the objectionably intrusive advertising procedures for membership and donations, all contributing to a dilution of its role as an independent evaluator; (iii) its use of an invalid synthesis process in assembling sub-evaluations, which has spoilt many of its major product evaluations. (More on the last point below.) There are other weaknesses of note, because they display poor analytic thinking, e.g., in the procedure they recommend for purchasing a car. (Further details in the forthcoming 5th edition of ***Evaluation Thesaurus***.)

[11] All too often, they are confused with wants assessments and done by survey of the interested parties. Such surveys have a place, but it is second place to the use of approaches such as dysfunction investigation and fault tree analysis.

XI. In sum, evaluation is a discipline which has been extensively developed both within and outside the traditional disciplines and has exactly the same logic in both situations. Its validity is obvious within—and hence as great as that of—other disciplines, and equally obvious in many practical applications such as product evaluation; and it is immune to the epistemological nihilism of the constructivist for the same reasons that other disciplines are immune. While the logical structure of evaluative claims is one step more complex than that of observational claims, their logic is not as complex as the logic of explanatory claims, and evaluative claims are usually easier to establish beyond reasonable doubt than most theoretical claims. What has *not* been developed until very recently is the core discipline of evaluation, because of a ban by the thought police; only in this respect does evaluation differ strikingly from the other transdisciplines.

CONSEQUENCES OF THE TRANSDISCIPLINARY VIEW

Having laid out some of the key theses of the transdisciplinary view of evaluation, we can turn to some of their consequences. First, three examples of consequences for our work that follow from the transdisciplinary view of the *general* nature of evaluation.

(i) To begin with, and perhaps most importantly, the TDV has a major feature that is intimately connected with our topic of accountability. This feature is not one of the theses just listed but it is a consequence of some work in the core discipline which consists in looking into the assumptions of the usual program evaluation procedure of measuring success against the goals of the program. Since the justification for doing this, outside the narrow confines of some program monitors' job descriptions, is that the goals of the program are meant to reflect the needs of its target population, and since the needs assessment on which that claim rests is frequently faulty in the first instance and just as frequently falsified by the constant ebbs and flows in the flux of programs available to this population, and the population's own changes in demographics and education, it follows that program evaluation usually has to go to checking or re-running the needs assessment in order to verify its findings.

Once this is realized, then one sees that any reference to the goals, while of some interest for formative evaluation, is essentially irrelevant to summative evaluation. So the TDV approach to program evaluation is a *consumer-oriented approach* intended to replace the *management-oriented approach* that dominated the early decades of development in program evaluation. That is, it focuses on *comparative cost-effectiveness in meeting the needs of consumers* (within the constraints of ethics) rather than being *on-time, on-budget, on-target,* i.e., meeting the goals of the program manager.

A review of the history of program evaluation lends further support to this point of view, for it shows that the original goal-focused approaches had to be modified in about seven ways to meet reasonable requirements coming from the audiences for evaluations.[12] Those modification also sum to a shift from using management goals as the key criterion of success to using 'consumer' (specifically, impactee) needs. That shift does not entail that there is something illicit about the manager's need to know whether a program is meeting its goals; good management requires good formative evaluation, and good formative evaluation will cover this point—but not only this point. The conclusion then, is simply that the demand that evaluators determine the extent to which a program has met its goals is not the only or even the primary legitimate demand on evaluation.

This general orientation of the TDV, if indeed it can be shown to follow from logical and pragmatic analysis of the concept of evaluation, connects it strongly with the intuitive feeling that there is a need for stressing accountability *to the citizenry* not just to managers. This feature is part of the reason that I felt it appropriate to bring in this theory as part of the "foundations of educational accountability" referred to in the subject of this paper. But it also has immediate consequence for our two school evaluation projects.

For it raises the question whether their validation procedures do in fact provide the connection with consumer needs that this model of evaluation suggests they should. My inclination is to say that they have to some degree fallen into the same trap as several models of evaluation, e.g., the transactional model and the negotiation model. That is, they get feedback from the actively and visibly involved communities of stakeholders—e.g., school administrators, staff, and parent groups—but spend less time getting input from the eventual consumers—employers, taxpayers, alums (and current) students, social change agents. Like the usual contract negotiations between the union and the school board, there's no expert representation of the students' interests, so we negotiate away the budget for materials—new texts, workbooks, computers—in favor of salary increases. This may not be so, but it is something to consider; and the same applies to administrator evaluation models.

8. School and administrator evaluation models should not be more influenced by input from visible and politically active stakeholders than by the interests of the ultimate consumers (students, taxpayers, employers, citizens).

In practice, this means making efforts to incorporate representatives of, or specialists in the effects on, these relatively invisible groups, in any review panels.

The second consequence of consumer orientation that we should consider applies at the metalevel, i.e., to evaluation of evaluations, here called meta-evaluation. The most important feature of school evaluation reports is validity; but the next most important is comprehensibility (to *all* of the intended audiences), and preferably readability. We need to approach this issue with just as much care as the validity, because it's the rock on which much indicator research has foundered. Good student report cards are easily understood by parents; sophisticated ones have often been rightly and bitterly attacked. We should make sure we avoid those attacks, even at the cost of some validity. Validity is not like pregnancy, and the emergence of 26 Offices of the

---

[12] The modifications were required in order to cope with the necessity for dealing with the following factors, as requested by clients and audiences: criticism of the goals of the program; the importance of side-effects and of shortfalls/overruns; the relevance of absolute values such as ethics and the law; the need for cost analysis and for comparisons; and the issue of generalizability. The dialog is set out in some detail in "Hard-Won Lessons in Program Evaluation: the 31 Theses" which constitutes the Summer, 1993 issue of ***New Directions in Program Evaluation*** (Jossey-Bass for the American Evaluation Association).

Inspector General at the federal level stands as a monument to the academic worship of what they took to be *absolute* validity when in fact it was a *degree* of validity associated with *fatal disutility*. We must not make the same mistake:

9. Readable school evaluation reports are as important as valid ones, and testing for comprehensibility/readability should be conducted before any reports are released.

(ii) The second major feature of the TDV is its "*wide-angle" emphasis*—the stress on evaluation as the broadest of all disciplines, and as one of the most fundamental of all disciplines. This has two implications: methodology transfer, and content overlap. On the first point, *every* evaluator should look at the way that Olympic dives are scored, or jockey's appeals settled, or Japanese students selected for college entry and graduation, as *part of their business*.

There are many cases, some already cited, where we have buried our head in the sands of our specialty and failed to notice what would have saved us from great loss of quality and wastage of resources in our own area. One example mentioned earlier were the errors in product evaluation even as the leading product evaluation institution does it. Another category of errors referred to in passing concerns the procedures used to evaluate proposals throughout the federal government and in most other contexts—the scoring scale which allocates percentage points to criteria of merit. This approach violates the elementary principle, established many years ago in the research field of personnel evaluation, that one must distinguish "compensatory" criteria from standalone criteria (the technical term for the latter is "multiple-cutoff"). This error has led to the misaward of millions and perhaps billions of dollars in contracts.

10. At this stage in the development of evaluation as a discipline, there is something important to be learnt—for good or for bad—from nearly every well-established process of evaluation in an autonomous field and much to be learnt from intradisciplinary evaluation in fields other than one's own.

Apart from picking up methodological hints, there is also the content-overlap point. The most obvious application to our present tasks is the need to ensure that the evaluation of schools is absolutely committed to checking on the ethics of school processes such as discipline and punishment, and the validity of school personnel evaluation procedures. Ethical appraisal and personnel evaluation are as much part of program or organizational evaluation as is cost analysis. Similarly, skill in teacher evaluation is an essential element to be rated in rating administrators (where the administrator has some responsibility for teachers—in other cases, skill in the evaluation of support personnel may be required). Naturally, this means that the rater must be skilled in that branch of evaluation. We must therefore expect from our administrator evaluation project a component dealing with teacher evaluation. Analogously, the teacher evaluator must rate—and hence be competent in rating—the teacher's skill in student assessment as a key element in teacher evaluation.

It is worth remembering that skill-reduction is a great attraction about compartmentalizing evaluation. I have seen efforts at job analysis for the teacher which omitted any reference to subject matter knowledge and I was not surprised to discover that most of the input came from school administrators. Naturally, the high school principal is not anxious to identify as essential, a skill they cannot possess in the requisite areas for all the teachers they have to evaluate. The substitution of method for content is not just a creation of professors of education. The transdisciplinary view, however, denies the validity of strong compartmentalization on the logical and commonsensical grounds that the job of evaluation takes whatever skills it takes to evaluate the dimensions of merit that must be assessed.

11. It is often true that efforts at one kind of evaluation, e.g., program evaluation, or at one sub-kind of evaluation, e.g., administrator evaluation (within personnel evaluation), must bring in reference to and use of another kind of evaluation—e.g., program evaluation may have to bring in personnel evaluation—or sub-kind of evaluation, and administrator evaluation may have to bring in teacher evaluation. It is inappropriate to publish models that either omit this cross-reference or fail to explain how it is to be done.

(iii) The third feature of the TDV to which it may be worth calling attention is its *technical emphasis*. It presents evaluation as a complex discipline, which includes a core discipline, autonomous applied areas of considerable complexity and sophistication, and the vast ranges of intradisciplinary evaluation. Mention of this quality is intended to offset the effects of the simple-minded but perennial identification by social scientists of the 'paradigm evaluation' as something like an expression of preference or a matter of taste. Such matters require no skill to produce—or to verify. For example, a statement like "I like this Chardonnay" or "I prefer excitement to security in a job" are not even evaluative (they are autobiographical statements), and involve no inference.

The college professors who were arguing for the value-free approach to science by using such examples in order to show that evaluations have no place in science should have used their own grading process as an example. It is certainly more typical of serious evaluation—performance assessment, by the way, not personnel assessment—and it might have suggested to them that the process relied on various assumptions about the validity of their test and their scoring of student work, assumptions that are often invalid for the procedures they were employing (e.g., grading on the curve, inappropriate use of multiple-choice tests, point constancy assumptions). They might then have learnt that these assumptions can be made valid by appropriate changes in the type of test and the marking procedures. These same value-free scientists were thus weakening their own science by engaging in invalid procedures of performance, proposal[13], and paper evaluation, while dismissing the field of systematic objective evaluation as a scientific absurdity. In short, treating evaluation as something like preferring was a costly error for their subject as well as their students.

There are five important points for our work here follow from the study of technical aspects of evaluation as it is conceived according to the TDV. I will not elaborate on them at length here since they are quite complex and treated fully in the ***Evaluation Thesaurus*** (Sage, 1991). The first point is the analog to the early work on scaling theory, which brought home to us the key differences between types of scale. Here, it is the crucial distinction between the four fundamental evaluation tasks: grading, ranking, scoring, and apportioning.[14] Every one of our projects needs to be clear that a different methodology is required for each of these tasks, and that it's expensive—and usually misleading or confusing—to use a methodology that covers tasks you do not have to perform. Should the school evaluations (for example) be aimed to rank or grade schools? For which audience? Is the request for ranking (or grading) legitimate or just idle curiosity? What will it cost to do each of the four tasks? Which is the most important, if we lack funds to do all? How do you decide that?

---

[13] The usual method of proposal evaluation simply allocates points for various dimensions of merit, a procedure which involves two or three fallacies. See Proposal Evaluation in ***Evaluation Thesaurus4***, op. cit.

[14] To call these fundamental is to say that none are reducible to a combination of the others. Incidentally, the first three can apply to both merit and worth, the fourth is worth-specific.

12. Evaluation designs that lead to ranking, grading, scoring, or apportioning are in general quite different, and the decision must be made early in our projects as to which is required and which is cost-feasible.

The second problem which emerges from increasing sophistication in the core discipline is perhaps the most serious problem in evaluation to have received essentially no attention at all in the evaluation literature. I call it *the synthesis problem*. It is the problem of finding a methodology for combining sub-evaluations into an overall evaluation. One approach, e.g., by Marv Alkin, is to reject the task as illicit and only report the sub-evaluations. This is an unsuccessful maneuver, in general, because: (i) the same problem is almost always to be found in the problem of reaching the sub-evaluations, each of which is typically itself a synthesis of evaluations as well as measurements; (ii) the client needs the synthesis, if it can be validly achieved, and lacks the skills to generate it validly in complex cases[15]. The usual way of doing it is via the "quantitative weight and sum" algorithm.[16] It is fairly easy to show that this is invalid, commonly involving large errors in complex cases (the ones where we need it most). I have devised an alternative, which appears to be valid but is not an algorithm and not as simple; it is a set of heuristics referred to as the "qualitative weight and sum" approach. Is the validity worth the complexity and partial decidability? For our work at the Center, I think we have to answer affirmatively, since the error size is often large. I suggest that in all our evaluation projects we should examine that matter, for it comes into every one we are doing.

13. The procedure for synthesizing sub-scores or sub-evaluations into an overall evaluation is technically sophisticated and needs to be given explicit attention in each project.

The third matter arising from technical considerations concerns a fundamental conceptual distinction, that between merit (or quality) and worth (or value). Merit/quality refer to something 'intrinsic' to whatever is being evaluated, measured against external standards such as the Personnel Standards (if we are looking at a system of teacher evaluation) or the Duties of the Teacher (if we are looking at a teacher). Worth/value refer to a quality that something has in relation to (and usually for) a organization or society of which it is part. Thus the worth of a high-school teacher depends not only on their merit but on the number of students that enroll in their course and on their salary. The worth of a researcher depends not only on their merit as a researcher but on the amount of grant revenue they can bring in. Notice that worth is much nearer the focus for an administrator than merit alone; it's also nearer the traditional notion of cost-effectiveness. Speaking for the TEMP project, it's pretty clear we simply have not paid enough attention to standards of worth—and I think this may be true of other projects. It may also be one reason why administrators do not pay as much attention to academic discussions of teacher evaluation as we evaluators would like them to. Note that the Personnel Standards cover at least some and perhaps all the relevant issue because of their inclusion of practicality considerations which of course include cost.

---

[15] It should be stressed that there are a minority of cases where it is better not to push for a synthesis, and some of them are important. One cannot prejudge the issue without some study of an evaluation task.

[16] Weight the criteria for importance on some scale, score the performances on a common scale (or normalize the scores), multiply the two for each candidate and sum the resulting products; the best candidate being the one with the highest score. There has been some primitive discussion of this in the measurement area under the awesome heading of "multi-attribute utility technology" (MAUT). The oversimplifications involved in that approach are discussed in *Evaluation Thesaurus, 4e*.

14. Evaluation of personnel and programs is not only a matter of determining merit but also worth; and the question of worth brings in matters of cost-effectiveness not usually seen as part of personnel evaluation.

The fourth issue that derives from technical investigation of the concepts in evaluation—easier now that the core discipline is opening up—concerns the difference between evaluation and meta-evaluation. Although the Personnel Standards are not *exactly* meta-evaluation, they are closely related to it, and this may illuminate the difference between them and the Duties-Based Model or any other mid-level model. In general, meta-evaluation is what administrators, especially high-level administrators, are directly concerned with: they usually have to be critical consumers of evaluation procedures rather than constructors of such procedures. One reason why meta-evaluation was late in achieving recognition was that we always had one way to do it that didn't seem to involve anything special—simply look at the evaluation and ask how you would have designed an evaluation for the task addressed, and compare that with the one you're looking at. We might call that "metaevaluation by reconsideration." But full-scale metaevaluation is something different. It is the evaluation of the evaluation under consideration *as a functional product*, and involves looking at the needs, effects, costs, and so on, just as one would for other products or programs. One small part of true metaevaluation involves looking at the design for technical adequacy—all that "metaevaluation by reconsideration" does. There are a number of checklists for doing metaevaluation, based on work by the GAO and others, and they represent the instrument of choice for that task. It seems plausible that we should look at each of our projects for coverage of meta-evaluation as well as evaluation tasks.

15. In general, thorough investigations of ways of evaluating entities should address the question how the client and audiences should evaluate the evaluations.

The last theoretically-based point of a technical kind to which I want to call attention concerns the relation between formative and summative evaluation. It applies to all of our projects with an evaluative component, but I'll illustrate it from teacher evaluation, where it may be more important than in any other application. There is a general prejudice towards thinking that evaluation for formative and summative purposes are quite different. Based on this view, teacher contracts have been written to allow *only* formative evaluation. But it is close to the logical heart of the matter to say that the only essential difference there should be between formative and summative evaluation is the different purposes to which they are put. Logically and methodologically they are essentially the same. In many contexts, the formative evaluation has a greater amount of detail, and sometimes it differs in the kind of detail it provides; but those differences are not always present—they are not part of the concepts. The formative evaluation *must* be at least a preview of good summative evaluation, because if it isn't that, it can't tell you how well you are doing overall, in the terms that define the job. And if it can't tell you that, then you can't tell where you need to improve and how much improvement is needed to reach the standard you wish to reach.

We have to distinguish previewing the actual kind of summative evaluation that will be done, which may be of very poor quality. While there is some pragmatic advantage to previewing that, perhaps by having a colleague role-play the principal, it does you little good in terms of professional development if it is going to be a shoddy evaluation. The essence of the point here is that good formative evaluation is only possible if based on the same evaluation logic as good summative evaluation.

I think part of the reason for confusion on this point is that people confuse formative evaluation with the provision of remedial or developmental advice. The evaluation is simply the *basis* in terms of which someone with appropriate remedial or developmental skills can provide advice. The evaluation locates us on the merit map: then, the expert on the territory can plot a route to guide us towards our destination. "Locating us" provides us with quite complex information; giving us the 'latitude and longitude' corresponds to telling us where we are on the merit profile. Location is done by celestial or satellite navigation; getting us through the mountain passes requires a different type of skill. Now of course there is some overlap. The formative evaluator, like the navigator, may easily generate some sensible suggestions as soon as the plot is located on the chart. But that's just a spin-off from the evaluation, it's not the task at which the navigator is the expert. And, of course, the navigator, once in possession of the fix, can easily determine how far we have to go, and in what overall direction we need to go—as the crow flies. In the mountains, that's not very helpful, and even topographic maps—if we had them—are a long way from replacing local knowledge.

16. We need to be clear that formative and summative evaluation are two sides of the same coin, logically and methodologically speaking; and we need to address directly the question how any system of evaluation that we recommend can be used to generate both formative and summative evaluation.

OTHER CONCLUSIONS

Descriptive vs. Normative Theories (and Analyses) Looking at our projects and the TDV, there is one general concern which comes immediately to mind. This concerns the nature of theories or analyses in applied evaluation fields like teacher evaluation. To begin with, there is some confusion between normative[17]—a.k.a. evaluative, sometimes prescriptive—and descriptive theories about teacher evaluation or administrator evaluation, etc. This confusion between can be found in the instructions we sometimes receive from Washington, requiring us to list the sources of evidence on which we are basing what is in fact a normative theory. One doesn't need extensive survey data or experimental evidence for a normative theory of statistical inference, nor for a normative theory of teacher evaluation. One only needs to understand the task, because what one is doing is working out how to do it. If the results are intended to have practical utility—by no means necessary, since some of the most important normative theories simply show the disutility of existing practices—then one must go to the field for demonstrate *the practical utility*. But one doesn't demonstrate practical utility by doing a survey, and one doesn't start analytical projects by gathering data from the field, one starts from the problem. The main procedure for confirming one's conclusions is not further data-gathering but analysis of a few more examples, as different as possible. Many problems in evaluation are clear enough the moment they are described by one person.

---

[17] A confusing term because it has built into it the hopeless attempt to reduce evaluation to statistical description (i.e., it incorporates the view that the scientific approach to evaluation is to treat it as reporting on phenomena in terms of the norms adopted by a group or exhibited by an ensemble). Evaluation does involve reference to norms but they are not empirical norms. rather evaluative norms i.e., standards of merit (not some group's standards of merit, but the correct standards of merit as near as we can judge). The proper use of the term "normative data" would be in reference to data reported in terms of deviation from some measure of central tendency, e.g., sigmas.

Now I suspect that this problem is not limited to Washington. It looks to me as if the self-assessment project, which is clearly descriptive, seriously needs a normative element. Otherwise, it will only be telling us how teachers *do* assess themselves, whether or not it's sensible or valid to do so. No normative element means no way to identify progress. I wonder, too, about the title of the "Evaluation Theory Development Project." In fact, it is simply a descriptive study of common practices in teacher evaluation, assessed against the Professional Standards; a complement to the kind of evaluation of mid-level teacher models we are doing in the TEMP project. This is not exactly the development of theory, something about which the project director knows as much as anyone in the business of evaluation. Hence I am not sure we're getting the most important contribution to personnel evaluation theory of which he's capable. (But we *are* getting something valuable, because his empirical basis is much larger than ours and the standards he's applying are quite different from ours—they are at the system level.)

On the other hand, the cross-cutting theory project seems to be going in the right direction, obviously selecting the right people for its conference and to provide its position papers!

Of great importance to us in the TEMP work is the difference between: (i) normative job analysis—which yields a duties statement; and (ii) descriptive job analysis—which can be done in two ways, either by time-sampling—which yields a report on what people in a job actually do, or by questionnaire—which yields a report on what they think they do, or think is most important about what they do (ETS), *neither of which is a statement of duties*. And only a statement of duties—so it is argued—can form a basis for personnel evaluation[18].

17. It is essential to distinguish between descriptive and normative theories and analyses, since different methodologies are required for each. A combination of them may be the best approach for several of our projects, but in combining them one needs to be clear about the difference between the methodologies required for the components, and the special methodology required for the combination.

Example of a purely critical finding To illustrate the point above about the payoff from normative theorizing sometimes having only critical effects, rather than "successful implementations", the small but potent "core discipline of evaluation" has already uncovered some essential flaws in established evaluation procedures. One example which bears on our concerns here is the discovery of a flaw in all currently institutionalized approaches to teacher evaluation, the flaw arising from the use of indicators obtained from the studies of teacher effectiveness as criteria for the evaluation of teaching. It is demonstrable although not obvious that this approach involves the same logical fallacy as "rationalized racism" i.e., discrimination in job selection against blacks or other minorities on the basis of, say, crime statistics which (let us say) in fact show that the incidence of crime amongst the group in question is higher than amongst Anglos. It may not even be obvious to some that rationalized racism is in fact racist; many others concede it to be racist in the ethical or legal sense, but think that not using it is one of the costs of combating racism. Analysis shows that it is no more and no less than an unscientific approach and increases the number of errors made and the money wasted; it is racist because it is unjustified discrimination on the basis of race[19].

---

[18] Of course, a statement of duties for e.g., teachers, is not going to be completely unlike a statement of what teachers think they do and/or think is most important about what they do, so the duties list has been improved as a result of looking at various job analyses. Note: for licensure, there is a better case for using job analyses than for any later personnel decisions, because of the numbers involved and the doubtful feasibility of more serious alternatives.

Applying this to our own activities, it raises the question whether any of our teacher or administrator or support personnel evaluation projects make this error, the error of using indicators derived from empirical or theoretical studies of best practice. They cannot afford even one such indicator since even one will contaminate use of the instrument to an unknowable extent.

18. An example of invalid normative methodology occurs in the use of correlation-validated indicators of merit in personnel evaluation. Unless no direct approach is possible because of constraints on time or evidence, any use of such indicators is invalid, and we need to examine our models for this error.

Levels of Standards A key issue concerns the extent of (i) coverage, and (ii) generalizability of various proposed standards in evaluation, particularly personnel evaluation. Pat Cross has recently argued that the existence of different paradigms for instruction in different areas of the curriculum shows that evaluation standards should be different as between those areas. In one sense, this is true. In another, such differences would be evidence of inequity. My present thinking is that there are three key levels at which standards can be set, that they cannot be completely separated, and that the third level has not received adequate attention by any of our projects. The three levels are, from the most general to the particular:

(i) the meta-standards level of the Standards for Personnel Evaluation. These set standards for *systems of teacher evaluation* rather than standards for direct evaluation of teachers, and I'll refer to them from now on as "system standards." While they are too general to be useful for carrying out (as opposed to setting up) teacher evaluation, they have the merit of generality in that they apply equally well to setting up standards for administrator and support personnel evaluation.

(ii) the mid-level standards, such as those set by research-based models of teacher evaluation, or by the duties of the teacher. These define the *generic* job of teaching—they define the profession—so I'll call them the *professional* standards, although there's some ambiguity in the term, and a better though less catchy term would be 'job-generic'. Although they *guide* the specific task of evaluating a particular teacher at a particular school, they are not enough to get that job done, for which we need a third level of standards. The mid-level standards are of no use for the evaluation of systems of teacher evaluation, *except* that they must be met if the system is to be valid. That is, they unpack a floating variable in the system standards, just as the third level of standards unpacks several floating variables (e.g., terms like "appropriate," "adequate") that occur in the mid-level standards.

(iii) The third kind of standards are the most specific, in fact specific to the particular job a teacher has to do—for example as a secondary math teacher, dealing mainly with advanced placement classes, in a parochial school. They might be called the *job-specific* standards, because they are related to what we call the "job description" and much more specific than the vocational standards. These standards tell us what kind of subject matter is to be covered at what level of difficulty. In many cases, it's arguable that they are too specific to be put into words, and trained judgment comes in. It's notable that both the Praxis effort and the National Board project are committed to filling in standards at this level.

---

[19] There are obviously some cases of justified discrimination on the basis of race, e.g., in hiring paid subjects for research on the incidence of sickle-cell anemia in Mediterranean peoples. But many seductive examples, e.g., hiring only blacks with an urban background for couselors to at-risk black youth in the inner city, do not hold up under pressure. An interesting borderline case is whether it's justified for the Navajo tribe to hire only Navajos for positions in the Navajo tribal police. Is this like hiring only Mormons for high offices in the Church of Latter Day Saints, or is it like hiring only black police officers for patrol duties in Harlem?

This is the level of standards that is normally used by the principal making a classroom visit, or in our "expert science teacher" project (where the identification of experts was done at widely separated sites by people who had no common training and no standards from a higher level in hand.

Dealing with this third level of standards is something I think we must address soon, and dealing with it will immediately mean trying to work out in more detail the relation between the Personnel Standards and the Duties (or other mid-level standards) and the (largely implicit) standards involved at the coal face.

19. It is important to separate three levels of standards with which the Center should be concerned: the level of system standards, the level of professional standards, and the level of job standards, the last being almost entirely implicit in our work at this point.

LAST WORD

There are many more topics I could have addressed and thus tried to relate our projects to each other and to those Great Standards In The Sky against which we will all be judged. But perhaps enough has been said to indicate that the effort to bring theoretical/conceptual considerations to bear on our work can lead to some useful suggestions.

APPENDIX D

POLITICS OF TEACHER EVALUATION

Gene V Glass
Arizona State University

Barbara A. Martinez
California State University-Los Angeles

POLITICS OF TEACHER EVALUATION

Gene V. Glass
Arizona State University

Barbara A. Martinez
California State University-Los Angeles

In this paper, we sketch the current practice of teacher evaluation in the U.S. from our ordinary experience. We present a fragment of a case study that reveals how teachers--elementary school (K-8) teachers in this instance--view the experience of being evaluated, and how feelings of illegitimacy flow from the experience. Finally, we indicate certain considerations from political theory that bear on the problems of improving teacher evaluation.

THE CURRENT STATE OF THINGS

Elementary school teachers are evaluated differently from how secondary school teachers are evaluated; and secondary school teachers are evaluated differently from college professors, who further underline the differences between themselves and their public school colleagues by not even wishing to be called teachers. Nothing seems to account for these differences so clearly as does what we might loosely refer to as the politics of evaluation. We often learn something interesting about the organization and politics of education when we contrast how it is pursued at its different levels.

College professors are usually evaluated by their peers and superiors yearly for raises and less often for promotion; but in spite of what might be claimed (by the president of the college when addressing parents, for an instance), they are seldom evaluated qua teachers. It is common today for students to fill out simple forms, rating scales, at the end of a semester: "Instructor was organized," "Instructor knew the subject," "Instructor graded fairly." Typically, these student ratings count for little. The better the university, the less teaching is weighed in the balance that sways toward research and publication; and most colleges aspire to be like the better universities. Extraordinarily bad student ratings will be used to terminate an untenured faculty member if that person's research is poor; but the administration will swallow bad ratings when a strong researcher receives them.

Secondary school teachers are evaluated sporadically. Peer evaluation is non-existent in America's schools, and administrators seldom venture into a high-school teacher's classroom. Their presence would be viewed with suspicion by the teacher; the legitimacy of their place there would be questioned (silently or behind closed doors if not publicly). Administrators appear to concede that secondary school teaching involves specialized knowledge (of chemistry or mathematics), and that a specialist may be needed to recognize good teaching.

Elementary school teachers are treated substantially differently. Principals visit their class once or more each year. Indeed, principals regard these visits as a responsibility of their position. Teaching is observed, occasionally a check-list is filled in; lesson plans may be inspected. Much is written in school personnel manuals about evaluating teachers on the basis of their students' achievement test scores; but the threat is an empty one that nonetheless has the power to shape instructional styles and choice of content (Glass, 1990). The legitimacy of the elementary school principal's presence in the classroom for the purpose of evaluating the teacher is less likely to be questioned openly. Everyone knows elementary school subject matter after all, and an instructional leader is an expert in the general techniques of effective instruction, or so it is widely believed.

Each level is distinguished by a different balance of teachers' professional autonomy on the one hand, and the exercise of administrative authority in a democratic bureaucracy on the other. At the elementary level, professional autonomy is difficult to discern and administrators are seen fulfilling the dictates of the duly elected school board to insure that teachers are competently delivering instruction to the students. At the secondary school level, teachers enjoy more autonomy to structure their classes and curriculum as they judge appropriate; administrative authority is exercised seldom and usually only in crises. College teachers enjoy autonomy granted by a three hundred-year tradition of academic freedom; no administrator dares to cross the threshold of the lecture hall.

Some have sought to reform teacher evaluation by attempting to alter the balance between these two forces. Art Wise and Tamara Gendler (1990), in *The New Handbook of Teacher Evaluation*, distinguished seven purposes and separate functions of teacher evaluation: preservice, selection, certification, "beginning," tenure, merit and school improvement. There is much that can be said about the politics of each of these separate phases of teacher evaluation (most of which would center on the politics of higher education and of the labor movement). Wise, for example, focuses on licensure and recommends a state board licensure system for teachers like that for physicians and lawyers. Such licensing might confer prestige on the profession and with prestige may come autonomy. But one might wonder whether medicine is well served by doctors or justice by lawyers. John McNeil (1981), in the *Handbook of Teacher Evaluation*, acknowledged the deep conflicts that surround this phenomenon in the school, but recommended new forms that scarcely differ from established practice and that fail to separate incompatible purposes for evaluating teachers. Armiger (1981), of the New Jersey Education Association, recommended guidelines for teacher evaluation that would give teachers more power within a system still "owned" by the bureaucracy.

It is our contention that the problems with teacher evaluation do not stem principally from the conflict between professional autonomy and bureaucracy (although these forces are apparent, they can not be changed without changing the political context), but from the perceived illegitimacy of the democratic bureaucracy in which the evaluation is embedded. Our argument will benefit from a portrayal of how teacher evaluation is experienced in the work lives of teachers.

TEACHER EVALUATION IN NOCAM

Nocam Elementary School is a K-8 school of 800 students located in the heart of a city of over two million people. Its attendance area is about a third Anglo and half Hispanic, with a smattering of children of many different cultures. Nocam is one of three elementary schools in the school district. It has a full-time principal and is closely linked to the Superintendent's office through the efforts of a curriculum specialist who has assisted Nocam is a major overhaul of its language arts curriculum. "Whole language" teaching, cooperative learning and non-graded organization have come to Nocam. In the course of pursuing a larger study focused on school reform, the second author conducted numerous interviews with Nocam teachers and administrators and observed classes, board meetings and teachers meetings. On the following pages, Nocam teachers speak of the way in which their work is evaluated by their superiors.

> District policy mandated that all teachers be evaluated once a year, despite the fact that there was no merit system of pay. Teacher evaluations were conducted by the district personnel director, except in the case of "new" teachers. "New" teachers, those who had been employed by the district for less than three years, were evaluated by their immediate supervisor, the school principal. Both new and veteran Nocam teachers viewed the evaluation process as "a joke," regardless of who the evaluator was. As one teacher explained, " . . . some man is going to come into my classroom, who has never been in my classroom all year and evaluate me on how good a teacher I am, by [observing] a twenty minute lesson [and]checking things off? That's impossible. I couldn't evaluate my students that way."
>
> Teachers generally shared this opinion, " . . . it's a scheduled appointment, they will be in your room at 10:30 and you have to have the handbook and the detention notices and the homework notices . . . they want to see homework, they want to see discipline records and it has to all be clearly posted, your discipline plan and everything . . . "
>
> Few of the veteran teachers were intimidated by the evaluation process; many, however, found the process coercive and demeaning. As one teacher explained, "I never see that personnel director except when he comes into make an appointment to do the observation . . . and you can't talk to him then. And then the next time that I see him is when he is handing me back my evaluation. And the thing is so arbitrary . . . it's 'you're an A teacher, you're a B teacher, you're a C teacher, and you fail.' I don't need to know if I'm an A teacher or a B teacher, I don't care. I care about whether or not I'm improving. They don't have enough respect for me . . . the amount of work that I have done and the amount of dollars that I have put in . . . to give me something that would actually help me improve. Instead they give me something that makes me work first of all to put something together for [the evaluator] to keep, then they want to evaluate my classes."
>
> Few teachers found the evaluation process informative or instructive. Teachers complained that the process failed to provide them with any insights as to how to improve their teaching. Frustrated by the procedure, teachers did not feel that the exercise was meant to help them improve their teaching, " . . . it's not meant to

improve [teaching] although that's the letter of the law, that the teacher evaluation systems are to improve teachers . . . to improve instruction. But it does not do that. In fact if it does anything I think it doesn't improve my instruction because I'm ticked for two days before I have to do it and I'm ticked for two days after I have to do it."

"New" teachers were a bit more anxious about the process, perhaps because they were evaluated by their principals. Like the veteran teachers, however, new teachers found the evaluation process much more burdensome than helpful. The evaluation experience described by Ms. Clark, a "new" K-3 Project teacher, was not unusual: "[The principal] kept saying ahead of time 'don't worry, I really need to see what's going on in your class'. I thought okay. And it was a time when I thought it was a good lesson, except for one kid, and this kid has been documented sexually abused, well he pooped in his pants during the lesson. The principal got hysterical with me, he was like 'I spent forty minutes in this room and I've seen nothing of value happening,' and he left. That was my first evaluation . . . . But I decided that I was going to talk to the principal [about my evaluation] and I said [to him], 'I'm a first year teacher and you can't just walk in my room and spend almost an hour and tell me nothing of value is going on. I'm not leaving until you tell me what good you saw.' So he decided not to count that [evaluation] and to do it again. So for the next one I prepared the kids . . ." Ms. Anderson was coached by her fellow teachers as to what this principal liked to see, the curriculum he preferred, and the practices he approved of. With this information in mind both she and her students practiced what they would do for the next evaluation: " . . . we practiced, we rehearsed what we would do when [the principal] walked in and I told the kids if we did it right, we got a surprise. And [other teachers] took the worst kids [to their classes that day] so that it wouldn't be bad. "It was awful . . . like the kids did these little work-sheets and they sat there and we had practiced what the work-sheets would look like, so they sat there and did them without talking. It was so awful . . . it was really hollow . . . we played the game. [The principal] told me I did a good job and I thought [to myself] 'you don't know anything.' I've learned part of the principal's game. I can do it when I have to, I've done it."

Stories about "putting on a show" for both evaluators and administrators were common among Nocam teachers: "The kids know how to act for the administrators. We bribe them [to act a certain way] when administrators are there. Then [the administrators] leave and we go back to our real way of working and of teaching." " . . . when the district people come into our class, I have to act a certain way, to put on a show . . . I train the kids to act the way the administrators expect them to act . . . even if the way they act [and the things we do] are not developmentally appropriate."

The administrators' ability to evaluate accurately teacher's performance was questioned by many teachers. Nocam teachers were of the opinion that their administrators did not really understand the pedagogical techniques nor the theoretical underpinnings of the techniques which were the basis of the K-3 Project, "I don't think they have any idea what [whole language or developmentally appropriate practice] look like. They think they are very supportive of whole language, but it's only as long as kids are sitting at their desks being really quiet." " . . . they don't understand how children learn and they come in and what can you tell them when you talk [with them] in the classroom for five minutes. Of course it looks like chaos . . . but learning is going on . . . you have to be there for a while to really get a grip on what is happening with the children . . . they don't understand [developmentally appropriate practice]. [The principal] evaluates you at a desk, a file cabinet between you and the reading group, he's not listening to the kinds of questions I'm asking the children or, you, know the communication skills going on. He's watching the behavior problems, and you are always going to have some. He's counting how many crayons the kids have on the floor." "They are not very supportive in the teaching methodology way, but more picky, and you have to do this and you have to teach from this book and you have to cover so much and you try to slip your own things in between without getting caught . . . You understand that [administrators] don't know much about [teaching] and you try to take it with a grain of salt." " . . . there are all kinds of politics on why we are getting raked over the coals for this and that, but if a principal does not understand what you are doing in the classroom before the evaluation starts, or if they don't agree with what you are doing, how can they evaluate you fairly . . . ?"

In addition to their not being very well grounded in the nontraditional models used by the K-3 Project teachers, many Nocam teachers were of the opinion that administrators didn't spend enough time in classrooms to accurately assess their teaching ability: " . . . they come in and what can you tell them when you talk [with them] in the classroom for five minutes . . . you have to be there for a while to really get a grip on what is happening with the children . . . " " . . . I don't really think he has a perfect understanding of what it is . . . of what exactly it is, because he doesn't come in our classrooms and hang out for an hour or two for a few days a year."

Nocam teachers also had concerns about the evaluators' ability, or lack thereof, to provide them with practical guidance and relevant assistance: "[The administrators] can't sit and discuss whole language theory with you . . . if they can't discuss the concepts with you how can they tell if what you are doing is right or wrong, and how can they help you improve upon it?" " . . . sometimes the things that they ask us to do don't particularly correspond with what we are trying to accomplish [in terms of teaching]." ". . . they don't understand how children learn . . . they don't understand developmentally appropriate practice . . .

I can't get much guidance from them . . ." "[The principal] says 'what can I do to help you,' but I feel that there is nothing he can do to help me because he doesn't have any knowledge to give me . . ." "I know they are busy [but] they need to spend more time in the classrooms with us. They come in once in a while, they make me nervous because they don't come in often enough for me to feel like they are friends, like I can ask them for help . . . I don't even know if they know what I'm doing in here." " . . . They might be supportive when you are explaining [the methodology] to them, the cooperative learning and the learning centers, but then they come into our rooms and see the movement and it's 'wait a minute, you didn't say kids were going to be talking to each other and moving around the classroom,' you said 'cooperative learning'."

Nocam teachers lacked guidance and direction. Even their direct supervisors--the school principals--lacked the appropriate pedagogical theory, and therefore were not a source of assistance or guidance. As one teacher explained: "I've really had to depend just on myself with all this. It's like I've been left out on this island, all alone . . . no guidance, no support, no validation . . . it's been pretty much a sink or swim situation . . . I still don't know which I'm doing . . . "

Nocam principals viewed their role primarily as that of "facilitator." Although these administrators encouraged their kindergarten through third grade faculty to "use whole language, cooperative learning, and developmentally appropriate practices," neither they nor the district superintendent provided teachers with concrete suggestions for implementing the techniques or improving their teaching. This task, according to the principals, was "left to the experts," who were brought in to provide in-service training throughout the school year.

The K-3 Project teachers managed to convince the district administration that it was unfair for teachers to be evaluated on their use of traditional teaching techniques when the K-3 Project relied so heavily on the use of nontraditional approaches. After three years of requesting, and largely in response to the requests of the K-3 Project coordinator, Nocam teachers were given a choice of being evaluated based on the standards of the "traditional teacher evaluation" instrument or based on the standards of a "whole language evaluation" instrument, recently developed by the district administration.

Both instruments assessed the same general categories of performance: classroom management, communication skills, instructional capabilities and materials, planning and organizational skills, compliance with school policies, and professional qualities. The whole language instrument, however, was much more extensive than the traditional instrument. The traditional evaluation instrument contained a total of five criteria per category which teachers could "exceed," "meet," or against which they could be judged "average" or "failed to meet." The whole language instrument contained twenty different criteria per category, which teachers could "exceed," "meet," or with respect to which they could be found "adequate" or "inadequate."

Technically, teachers had a choice in the evaluation matter; practically they did not. Neither the district evaluator nor the principals were adept in using the nontraditional instrument. As a result, though a fair number of teachers requested that the new instrument be used, only one teacher was actually assessed with it. The process was described by the teacher as "a disaster."

According to this teacher, the district evaluator did not have the new forms in his evaluation package when he came to evaluate her, nor did he know what was on the forms. The teacher had to supply the evaluator with the new forms. The evaluator didn't understand the stated criteria; he had the teacher explain to him how the new criteria related to the old, so he would know what to look for during the evaluation session. The teacher described thus: "It was just a disaster . . . [The district evaluator] couldn't sit and discuss whole language theory with me . . . he doesn't know a thing about developmentally appropriate practice . . . and cooperative learning . . . forget it . . . he [kept] looking for my assertive discipline program. That's not my priority . . . I had to explain the entire process to him, what to look for, what was appropriate and why. I'm sure he learned a lot, if he paid any attention, but for him to evaluate me, what a joke."

Whether they knew how to use the nontraditional evaluation instrument or not, the district principals avoided using it. The principal at one of the other two elementary schools in the district went as far as to tell his teachers that it was his choice which instrument was used, not theirs. And he chose to use only the traditional instrument.

One teacher decided to "check this out with the district." She was informed that the teachers did indeed have the right to choose. When she shared this information with her principal he "got upset at me for questioning his authority and he told me that he was going to talk with the district people himself. In the meantime, he used the traditional instrument to evaluate me."

The Nocam district superintendent maintained that the district allowed teachers "the freedom to use what they think is appropriate" in teaching their classes. He also believed that someone needed "to make sure that what they feel is appropriate is in line with . . . our curriculum philosophy . . . [that it is] highly matched to what we expect kids to be tested on." A number of tools were developed to assist Nocam administrators in monitoring curriculum and instruction.

The Nocam case makes one thing clear. Even at the elementary school level where it might be expected that administrative evaluation is most defensible, it is viewed by teachers as illegitimate. Principals are seen as uninformed about curriculum and unable to spend the time to understand the circumstances of the class in such a way that they could help improve it. Bureaucratic evaluation of teaching at the secondary and college levels is seen as an affront to professional autonomy and as being even less legitimate than elementary school teacher evaluation.

It cannot be argued that what was seen in Nocam is somehow an outgrowth of the special circumstances of poverty or ethnic minority culture. Similar experiences are widely spread in the American educational system and, perhaps, elsewhere (Beery, 1992).

LEGITIMIZING TEACHER EVALUATION

It is our contention that the principal problem with teacher evaluation is that it is viewed as lacking legitimacy by the persons who are the object of the evaluation, the teachers themselves.

Consequences of Loss of Legitimacy

Teacher evaluation viewed as illegitimate by teachers themselves generates nothing but dissembling, passivity and feelings of alienation and powerlessness (Glass, 1990). School boards through administrators have a legitimate interest in how instruction is conducted, but it is not an overriding interest, nor does it follow that their interest is served by direct participation of the principal in the evaluation of teachers.

Where legitimacy is lacking, one can expect little more than passive compliance. Is it a matter of concern that an evaluation system is imposed from the administrative hierarchy and not seen as legitimate by the teachers who are being evaluated? One line of argument answers "No." Suppose that the system imposed is so comprehensive and well designed that it encompasses most of what teachers should be expected to perform. One might argue then that it is irrelevant whether the teachers "like" what it imposes, since if they conform to its vision of what a good teacher is, they will ipso facto teach well. This argument is similar to questions debated under the topic of "teaching to the test" in educational assessment. Some maintain that if a test is good enough, then teaching to it will only result in good education. Similarly, if a teacher evaluation system is good enough (i.e., defines a good model of what a teacher is to be), then complying with it--even if that compliance is "false" or pretended in some sense--will result in the teacher being a good teacher.

The difficulty with the "teaching to the test" argument is that the kinds of test generally used in assessments are rather pale reflections of a good education. Likewise, many of the bases of teacher evaluation systems are weak or impoverished models of what good teaching is. Checklists of teaching acts or "elements" reduce teaching to a few general principles of instruction, and divert attention from concerns of curriculum. A teacher can be a good teacher under such surveillance while teaching shallow or false knowledge. Some believe that this is little concern at elementary grades since "there is no discipline" (in the academic sense) at that level. "Elementary school teachers have no discipline, they just teach;" or "teaching is their discipline." Others are shocked to hear that the teaching of reading or language or mathematics is believed by some not to raise technical and intellectual questions as complex and sophisticated as the teaching of calculus to high-school students.

Ways of Seeking Legitimacy

Legitimacy can be bestowed in at least two ways: by appeals to widely accepted scientific or technological knowledge or through the appeal to the authority of legitimate political institutions or arrangements.

The attempt to legitimate the standard practice of teacher evaluation by appealing to science and technical-rationality fails for a couple of reasons. First, there is no widely respected science of teaching and learning. Common sense or practical and tacit knowledge of teaching usually succeed as well as systems that profess to be based on research. Second, most efforts to reform teacher evaluation start from an assumption that all parties with a direct interest in improved education share a consensus on what good education is. From false assumptions of consensus come technical-rational attempts to manage teachers. We begin from a different starting point. Schools are micro-political units where teachers, administrators, parents, students and even society far removed from the classroom seek to realize their interests. These interests often conflict. Without agreement on ends, mechanical and technical solutions fail. Third, school administrators who are vested with the authority to evaluate teachers as instructors generally lack knowledge of the subject matter being taught. Their role as evaluator consequently strikes the teacher as superficial, then illegitimate. (See Scriven, 1992, for a discussion of what limits subject matter specificity does and does not place on teacher evaluation.)

An example may help illustrate a nonspecialist's lack of subject matter knowledge can invalidate the type of evaluation that focuses on general acts of teaching. A junior high school English teacher is reviewing a lesson on nouns. He writes on the board "A noun is a Person, Place or Thing," leaving space beside each for examples. Turning to the class, he invites anyone who can illustrate a noun as the name of a person to step to the board and write. A student is congratulated for writing "singer" beside Person, as is a second student who writes "school" beside Place. The third volunteer writes "kitchen" beside Thing only to be told politely by the teacher that a kitchen is a place, not a thing. The checklist that the observer of this episode was filling out had categories only for the commonly identified important elements of teaching: previews lesson, clarifies goals, provides for active participation, reinforces correct responses, and the like. This teacher even scores points for correcting mistakes quickly. This kind of evaluation of this act of teaching misses the important point of what actually took place. One need not be a grammarian to sense that something is seriously wrong with this lesson. Of course a kitchen is a "thing," as in "I have remodeled my kitchen." And it may function as a place in other uses. Nor are Person, Place and Thing mutually exclusive and exhaustive categories. Where does the "unicorn" reside, grammatically speaking, and what about "truth"? The point is that this teacher is teaching a shallow or false point and this consideration should override all other questions. Indeed, his grasp of grammar has led him to confuse a student (probably more than one) and draw that student into a publicly embarrassing situation, where her valid understanding of language usage is labeled "wrong." Where in the evaluation of this teacher is it noted that the teacher has a responsibility to understand and continue to learn the subject being taught? Some will argue that it is impossible for a principal or a principal's deputy to know all the subject matter taught by all the teachers in the school. Indeed it is; and lacking that understanding, it is questionable to what degree the principal can serve as the guide for the teacher's efforts to become a better teacher.

If technical-rationality can not confer legitimacy on teacher evaluation, then it remains for political arrangements to do so. Modern political institutions are bureaucratic democracies, with one distinguished from the other by the balance of democracy and bureaucracy.

Appeals to the science of teaching or to technical-rational arguments about what ought to be taught and how can not hope to justify a particular form for teacher evaluation within the hierarchical bureaucracy of contemporary schools. Legitimacy for some form of teacher evaluation must be found in a new set of political arrangements that will be viewed as legitimate by the teachers themselves.

Seeking New Political Arrangements

Three evaluation theorists have addressed the systemic political problems that have led to the current state of teacher evaluation.

MacDonald. "Democratic evaluation," as envisioned by Barry MacDonald (1974), addresses the tension between power concentration and power diffusion in liberal democracies by opting for radical power diffusion. MacDonald focuses more on the role of the evaluator than on such aspects of the evaluation as the criteria, data and the like. He saw the evaluator as an information broker among interested parties. The evaluator stops short of making recommendations; information, ultimately owned by those from whom it is collected, is presented to those persons with legitimate interests in what is being evaluated. Decisions flow from some unspecified process of democratic discussion among interested parties. "MacDonald's evaluation approach intentionally includes diverse interests, allows people to represent their own interests, and is based on an idea of mutual consent" (House 1980, p.150). A direct, rather than a representative form of democracy, is being imagined by MacDonald. The limitations of direct democratic participation in complex, mass societies are obvious. However, one is casting a small net when the object of an evaluation is a teacher and a classroom. The range of interests to capture and bring to consensus is narrow. After all, juries reach consensus even when the stakes are high.

Strike. Kenneth Strike has pursued an examination of political forms and their relevance to education. He contrasts two quite different approaches to achieving democratic governance. The first is John Locke's legislative majoritarian democracy. Its operations are familiar to all; its assumptions are less obvious. Naturally free and equal humans are to be granted equal sovereignty, which is exercised by voting for representative government. With the consent of the governed, sovereignty is placed in the legislative body. Legislatures exercise sovereign control through hired managers who follow the policy direction set down by the legislative body.

Exigencies will call for clearer rules and policies; in time, rules will accrue and the modern bureaucratic democracy will emerge. Citizens may not regard every rule as legitimate, but the stability of the institution rests on a wide-spread belief in the legitimacy of the process by which the representatives are first chosen and then formulate the rules. "We can see a legislature as a means to vector interests more than as a means for making and judging the merits of practical argument. Majorities may be seen as formed more by a process of combining and reconciling interests than by a process that seeks the better argument." (Strike, 1993, p. 16). Current practice in teacher evaluation is embedded in this context of legislative bureaucratic democracy. It has failed to engender among teachers a belief in its legitimacy. Most writing to date about the politics of teacher evaluation have assumed no changes in the basic nature of democratic institutions and as a result offer suggestions that tinker with the balance of power between democratic bureaucracies and professions.

An opposing view of democratic institutions grows out of the attempts of Jurgen Habermas to justify liberal democracy. Habermas argues for the legitimacy of a communitarian democracy in which social norms are justified by uncoerced argument among equals in an ideal speech community. To Habermas, a social choice is "discursively redeemed" when it has the consensus of a community of citizens and that consensus was reached in open and undominated discourse. Argument--not votes--legitimates choices and actions for the good of the community.

Strike recognizes a Utopian character to Habermas's notion of the discursive redemption of policy choices in an ideal speech community. As a practical matter, sovereignty will have to be located in a representative body and conflicting interests "vectored" to a solution when consensus is impossible. But he can not back away from Habermas's ideas without trying to answer the question, "How might we make bureaucratic democratic institutions more Habermasian?"

Kemmis. While Strike may wish to cook the Habermasian omelet without breaking the Habermasian eggs, Stephen Kemmis does not hesitate to recommend the Habermasian ideal. To Kemmis, teacher evaluation would be one particular aspect of what he calls "emancipatory action research" (Carr Kemmis, 1986; Kemmis, 1993a; Kemmis, 1993b). "When schools--teachers, students, principals and others--are forced to change on the basis of outside evaluations and the crude coercive powers of the state, however, they frequently resist, passively if not actively. And that, it seems to me, just produces still further administrative demands for surveillance, regulation and control . . . I believe that the evaluation processes I have attempted to develop--as well as some of the practices associated with 'responsive,' 'illuminative,' and 'democratic' evaluation--did (and do) contribute to the development of less irrational, less unjust and less satisfying forms of social life. Though some of those perspectives have no particular inclination to justify themselves against the criteria of critical social theory or critical social science, in practice they do seem to offer increased opportunities for what Habermas describes as 'communicative action'--action oriented towards mutual understanding and unforced agreement . . . " (Kemmis, 1993, pp. 46-47).

CONCLUSION

Teachers view the evaluation to which they are subject as being illegitimate. They do not recognize the authority of those who perform the evaluation; they do not accept it as valid and defensible. Legitimacy can be conferred by democratizing the process of teacher evaluation, by removing it from the context of hierarchical bureaucracy in which it now resides, and by carrying it out in a new context. Some theorists offer justifications for this rearrangement of the politics of teacher evaluation. As yet, it is unclear how the reform would be played out in its essential details, e.g., who would participate in the evaluation of teachers, what information would be relevant and how it would be obtained, with whom authority would lie to call for an evaluation, and the like.

POSTSCRIPT ON CREATE AND ITS PROGRAMS

We have been asked to address how our paper bears on CREATE and its programs of research and development. We must confess to knowing less about CREATE's programs than we should. However, two prominent examples of CREATE's work are at hand: memos from the Teacher Evaluation Models Project, and a draft chapter entitled "Evaluation for Effective Teaching" by Carol A. Dwyer and Daniel L. Stufflebeam (a chapter in press in the Handbook of Educational Psychology). We do not necessarily take the Dwyer Stufflebeam paper as the official position of CREATE on teacher evaluation; Scriven's memos more clearly carry the CREATE imprimatur. But, Stufflebeam is the Director of CREATE, after all.

Scriven's (1991, 1992) duties-based evaluation speaks to the criteria for evaluating teachers and does not presume any particular political arrangements under which it is carried out. His approach is not overly technical--unlike bureaucratic approaches that attempt to grasp legitimacy by dint of scientism. The list of duties appears to have been put together with some sensitivity to how teachers view their role. In a system where teachers are empowered to define and participate equally in their own evaluation, it is not hard to imagine that Scriven's list of duties could be selected as a starting point.

The Dwyer and Stufflebeam manuscript is significant because of its scope; it is seventy single-spaced pages long. They reference 202 other published works. Our fifteen single spaced pages cite 20 references. The intersection of their and our references is a set with four elements: two Scrivens (both duties based), and one Glass (which we would cite whether relevant to the topic at hand or not). Admittedly, we should probably read a couple of Dwyer and Stufflebeam's references that we either missed or never got around to reading (e.g., the Darling-Hammond Wise Review of Educational Research piece on "teacher evaluation in the organizational context" and the paper by Milner in the Kappan entitled "Suppositional style and teacher evaluation"). But for all intents and purposes, the two papers have almost nothing in common. From our perspective, Dwyer and Stufflebeam accept the current political organization of teacher evaluation as a "given" and offer no critique of it. We see the current political arrangement as vitiating the potential good that teacher evaluation can contribute to education.

REFERENCES

Armiger, M. L. (1981). The political realities of teacher evaluation. In J. Millman (Ed.), *Handbook of Teacher Evaluation* (pp. 292-302). Beverly Hills, CA: Sage.

Beery, B. F. (1992). *Master teacher conceptions of relationships between teacher evaluation and excellence in teaching performance*. Doctoral dissertation. Tempe, AZ: Arizona State University.

Carr, W., & Kemmis, S. (1986). *Becoming critical: Education, knowledge and action research.* London: Falmer Press.

Glass, G. V. (1990). Using student test scores to evaluate teachers. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 229-240). Newbury Park, CA: Sage.

House, E. R. (1980). *Evaluating with validity.* Beverly Hills, CA: Sage.

Kemmis, S. (1993a). Action research and social movement: A challenge for policy research. *Education Policy Analysis Archives, 1*(1) (entire issue).

Kemmis, S. (1993b). Foucault, Habermas and evaluation. *Curriculum Studies, 1*(1), 33-52.

MacDonald, B. (1974). *Evaluation and the control of education.* Norwich, England: Centre for Applied Research in Education.

Martinez, B. A. (1993). *How educational reform is compromised: A critical investigation.* Doctoral dissertation. Tempe, AZ: Arizona State University.

McNeil, J. D. (1981). Politics of teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 272-291). Beverly Hills, CA: Sage.

Scriven, M. (October 1992). *Should teacher evaluation be subject-matter specific*? (TEMP Memo #10). Kalamazoo, MI: Center for Research on Educational Accountability Teacher Evaluation, Western Michigan University.

Scriven, M. (September, 1991). *Duties of the teacher* (Memo). Kalamazoo, MI: Center for Research on Educational Accountability Teacher Evaluation, Western Michigan University. (Also see Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education.*)

Strike, K. A. (in press). Professionalism, democracy and discursive communities: Normative reflections on restructuring. *American Educational Research Journal.*

Strike, K. A. (1990). Is teaching a profession? How would we know? *Journal of Personnel Evaluation in Education, 4*, 91-117.

Strike, K. A. The moral role of schooling in a liberal democratic society. *Review of Research in Education, 17.*

Strike, K. A. (1991). Humanizing education: Subjective and objective aspects. *Studies in Philosophy of Education, 11*(1), 17-30.

Strike, K. A. (1990). The ethics of educational evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 356-373). Newbury Park, CA: Sage.

Strike, K., & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 303-343). Beverly Hills, CA: Sage.

Wise, A., & Gendler, T. (1990). Governance issues in the evaluation of elementary and secondary school teachers. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 374-389). Newbury Park, CA: Sage.